
EvalML Documentation

Release 0.32.1

Alteryx, Inc.

Sep 10, 2021

CONTENTS

1	Install	3
2	Start	7
3	Tutorials	13
4	User Guide	43
5	API Reference	155
6	Release Notes	1307
	Python Module Index	1355
	Index	1359

EvalML is an AutoML library that builds, optimizes, and evaluates machine learning pipelines using domain-specific objective functions.

Combined with [Featuretools](#) and [Compose](#), EvalML can be used to create end-to-end supervised machine learning solutions.

INSTALL

EvalML is available for Python 3.7 and 3.8 with experimental support 3.9. It can be installed with pip or conda.

1.1 Time Series support with Facebook's Prophet

To support the `Prophet` time series estimator, be sure to install it as an extra requirement. Please note that this may take a few minutes. Prophet is currently only supported via pip installation in EvalML for Mac with CmdStan as a backend.

```
pip install evalml[prophet]
```

Another option for installing Prophet with CmdStan as a backend is to use `make installdeps-prophet`.

Note: In order to do this, you must have the EvalML repo cloned and you must be in the top level folder `<your_directory>/evalml/` to execute this command. This command will do the following: - Pip install `cmdstanpy==0.9.68` - Execute the `install_cmdstan.py` script found within your `site-packages/cmdstanpy` which builds `cmdstan` in your `site-packages`. - Install `Prophet==1.0.1` with the `CMDSTAN` and `STAN_BACKEND` environment variables set.

If the `site-packages` path is incorrect or you'd like to specify a different one, just run `make installdeps-prophet SITE_PACKAGES_DIR=<path_to_your_site_packages>`.

If you'd like to have more fine-tuned control over the installation steps for Prophet, such as specifying the backend, follow these steps:

For `CmdStanPy` as a backend: 1. `pip install cmdstanpy==0.9.68` 2. `python <path_to_installed_cmdstanpy>/install_cmdstan.py --dir <path_to_build_cmdstan>`
3. `CMDSTAN=<path_to_build_cmdstan>/cmdstan-2.27.0 STAN_BACKEND=CMDSTANPY pip install prophet==1.0.1`

For `PyStan` as a backend (`PyStan` is used by default): 1. `pip install prophet==1.0.1`

1.2 Pip with all dependencies

To install evalml with pip, run the following command:

```
pip install evalml
```

1.3 Pip with core dependencies

EvalML includes several optional dependencies. The `xgboost` and `catboost` packages support pipelines built around those modeling libraries. The `plotly` and `ipywidgets` packages support plotting functionality in automl searches. These dependencies are recommended, and are included with EvalML by default but are not required in order to install and use EvalML.

EvalML's core dependencies are listed in `core-requirements.txt` in the source code, and optional requirements are listed in `requirements.txt`.

To install EvalML with only the core required dependencies, download the EvalML source from [pypi](#) to access the requirements files. Then run the following:

```
pip install evalml --no-dependencies
pip install -r core-requirements.txt
```

1.3.1 Add-ons

You can install add-ons individually or all at once by running:

```
pip install evalml[complete]
```

Time Series Support

Add time series support with Facebook's Prophet

```
pip install evalml[prophet]
```

Please note that this may take a few minutes. Prophet is currently only supported via pip installation in EvalML.

Update checker

Receive automatic notifications of new EvalML releases

```
pip install evalml[update_checker]
```

1.4 Conda with all dependencies

To install evalml with conda run the following command:

```
conda install -c conda-forge evalml
```

1.5 Conda with core dependencies

To install evalml with only core dependencies run the following command:

```
conda install -c conda-forge evalml-core
```


1.6 Windows

Additionally, if you are using `pip` to install EvalML, it is recommended you first install the following packages using `conda`: * `numba` (needed for `shap` and prediction explanations). Install with `conda install -c conda-forge numba` * `graphviz` if you're using EvalML's plotting utilities. Install with `conda install -c conda-forge python-graphviz`

The `XGBoost` library may not be `pip`-installable in some Windows environments. If you are encountering installation issues, please try installing XGBoost from [Github](#) before installing EvalML or install `evalml` with `conda`.

1.7 Mac

In order to run on Mac, `LightGBM` requires the `OpenMP` library to be installed, which can be done with `HomeBrew` by running

```
brew install libomp
```

Additionally, `graphviz` can be installed by running

```
brew install graphviz
```

1.8 Python 3.9 support

`Evalml` can still be installed with `pip` in python 3.9 but note that `sktime`, one of our dependencies, will not be installed because that library does not yet support python 3.9. This means the `PolynomialDetrending` component will not be usable in python 3.9. You can try to install `sktime` [from source](#) in python 3.9 to use the `PolynomialDetrending` component but be warned that we only test it in python 3.7 and 3.8.

START

In this guide, we'll show how you can use EvalML to automatically find the best pipeline for predicting whether or not a credit card transaction is fraudulent. Along the way, we'll highlight EvalML's built-in tools and features for understanding and interacting with the search process.

```
[1]: import evalml
      from evalml import AutoMLSearch
      from evalml.utils import infer_feature_types
```

First, we load in the features and outcomes we want to use to train our model.

```
[2]: X, y = evalml.demos.load_fraud(n_rows=250)
```

```

      Number of Features
Boolean                1
Categorical             6
Numeric                5

Number of training examples: 250
Targets
False    88.40%
True     11.60%
Name: fraud, dtype: object
```

First, we will clean the data. Since EvalML accepts a pandas input, it can run type inference on this data directly. Since we'd like to change the types inferred by EvalML, we can use the `infer_feature_types` utility method. Here's what we're going to do with the following dataset:

- Reformat the `expiration_date` column so it reflects a more familiar date format.
- Cast the `lat` and `lng` columns from float to str.
- Use `infer_feature_types` to specify what types certain columns should be. For example, to avoid having the `provider` column be inferred as natural language text, we have specified it as a categorical column instead.

The `infer_feature_types` utility method takes a pandas or numpy input and converts it to a pandas dataframe with a [Woodwork](#) accessor, providing us with flexibility to cast the data as necessary.

```
[3]: X.ww['expiration_date'] = X['expiration_date'].apply(lambda x: '20{}-01-{}'.format(x.
      ↪split("/") [1], x.split("/") [0]))
      X = infer_feature_types(X, feature_types= {'store_id': 'categorical',
      'expiration_date': 'datetime',
      'lat': 'categorical',
      'lng': 'categorical',
      'provider': 'categorical'})

      X.ww
```

```
[3]:
```

	Physical Type	Logical Type	Semantic Tag(s)
Column			
card_id	int64	Integer	['numeric']
store_id	int64	Integer	['numeric']
datetime	datetime64[ns]	Datetime	[]
amount	int64	Integer	['numeric']
currency	string	Unknown	[]
customer_present	bool	Boolean	[]
expiration_date	datetime64[ns]	Datetime	[]
provider	category	Categorical	['category']
lat	float64	Double	['numeric']
lng	float64	Double	['numeric']
region	category	Categorical	['category']
country	category	Categorical	['category']

In order to validate the results of the pipeline creation and optimization process, we will save some of our data as a holdout set.

```
[4]: X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
    ↪problem_type='binary', test_size=.2)
```

Note: To provide data to EvalML, it is recommended that you initialize a woodwork accessor so that you control how EvalML will treat each feature, such as as a numeric feature, a categorical feature, a text feature or other type of feature. Consult the [the Woodwork project](#) for help on how to do this. Here, `split_data()` returns dataframes with woodwork accessors.

EvalML has many options to configure the pipeline search. At the minimum, we need to define an objective function. For simplicity, we will use the F1 score in this example. However, the real power of EvalML is in using domain-specific *objective functions* or *building your own*.

Below EvalML utilizes Bayesian optimization (EvalML's default optimizer) to search and find the best pipeline defined by the given objective.

EvalML provides a number of parameters to control the search process. `max_batches` is one of the parameters which controls the stopping criterion for the AutoML search. It indicates the maximum number of rounds of AutoML to evaluate, where each round may train and score a variable number of pipelines. In this example, `max_batches` is set to 1.

**** Graphing methods, like AutoMLSearch, on Jupyter Notebook and Jupyter Lab require [ipywidgets](#) to be installed.**

**** If graphing on Jupyter Lab, [jupyterlab-plotly](#) required. To download this, make sure you have [npm](#) installed.**

```
[5]: automl = AutoMLSearch(X_train=X_train, y_train=y_train,
    problem_type='binary', objective='f1',
    max_batches=1, verbose=True)
```

```
Generating pipelines to search over...
Removing columns ['currency'] because they are of 'Unknown' type
8 pipelines ready for search.
```

When we call `search()`, the search for the best pipeline will begin. There is no need to wrangle with missing data or categorical variables as EvalML includes various preprocessing steps (like imputation, one-hot encoding, feature selection) to ensure you're getting the best results. As long as your data is in a single table, EvalML can handle it. If not, you can reduce your data to a single table by utilizing [Featuretools](#) and its Entity Sets.

You can find more information on pipeline components and how to integrate your own custom pipelines into EvalML [here](#).

```
[6]: automl.search()
```

```
*****
* Beginning pipeline search *
*****
```

```
Optimizing for F1.
Greater score is better.
```

```
Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 9 pipelines.
Allowed model families: catboost, linear_model, decision_tree, extra_trees, random_
↳forest, xgboost, lightgbm
```

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...}
```

```
Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.000
```

```
*****
* Evaluating Batch Number 1 *
*****
```

```
Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler + Standard Scaler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.236
```

```
Decision Tree Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.199
```

```
Random Forest Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.403
```

```
LightGBM Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.279
```

```
Logistic Regression Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler + Standard Scaler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.197
```

```
XGBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.205
```

```
Extra Trees Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
```

```
  Starting cross validation
  Finished cross validation - mean F1: 0.193
```

```
CatBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + Oversampler:
```

(continues on next page)

(continued from previous page)

```

Starting cross validation
Finished cross validation - mean F1: 0.538

Search finished after 00:17
Best pipeline: CatBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + Oversampler
Best pipeline F1: 0.537892

```

If you would like to suppress stdout output, set `verbose=False`. This is also the default behavior for `AutoMLSearch` if `verbose` is not specified.

```
[7]: automl = AutoMLSearch(X_train=X_train, y_train=y_train,
                          problem_type='binary', objective='f1',
                          max_batches=1, verbose=False)

automl.search()
```

We also provide a *standalone* `search` method `<./generated/evalml.automl.search.html>` which does all of the above in a single line, and returns the `AutoMLSearch` instance and data check results. If there were data check errors, AutoML will not be run and no `AutoMLSearch` instance will be returned.

After the search is finished we can view all of the pipelines searched, ranked by score. Internally, EvalML performs cross validation to score the pipelines. If it notices a high variance across cross validation folds, it will warn you. EvalML also provides additional *data checks* to analyze your data to assist you in producing the best performing pipeline.

```
[8]: automl.rankings
```

	id	pipeline_name	search_order	\
0	8	CatBoost Classifier w/ Drop Columns Transforme...	8	
1	3	Random Forest Classifier w/ Drop Columns Trans...	3	
2	4	LightGBM Classifier w/ Drop Columns Transforme...	4	
3	1	Elastic Net Classifier w/ Drop Columns Transfo...	1	
4	6	XGBoost Classifier w/ Drop Columns Transformer...	6	
5	2	Decision Tree Classifier w/ Drop Columns Trans...	2	
6	5	Logistic Regression Classifier w/ Drop Columns...	5	
7	7	Extra Trees Classifier w/ Drop Columns Transfo...	7	
8	0	Mode Baseline Binary Classification Pipeline	0	

	mean_cv_score	standard_deviation_cv_score	validation_score	\
0	0.537892	0.201574	0.769231	
1	0.402694	0.040471	0.400000	
2	0.278620	0.074925	0.222222	
3	0.236111	0.122615	0.142857	
4	0.204545	0.186031	0.000000	
5	0.199216	0.075506	0.213333	
6	0.197114	0.063306	0.142857	
7	0.193023	0.167490	0.279070	
8	0.000000	0.000000	0.000000	

	percent_better_than_baseline	high_variance_cv	\
0	53.789174	False	
1	40.269360	False	
2	27.861953	False	
3	23.611111	False	
4	20.454545	False	
5	19.921569	False	
6	19.711400	False	

(continues on next page)

(continued from previous page)

```

7          19.302326          False
8          0.000000          False

                                parameters
0  {'Drop Columns Transformer': {'columns': ['cur...
1  {'Drop Columns Transformer': {'columns': ['cur...
2  {'Drop Columns Transformer': {'columns': ['cur...
3  {'Drop Columns Transformer': {'columns': ['cur...
4  {'Drop Columns Transformer': {'columns': ['cur...
5  {'Drop Columns Transformer': {'columns': ['cur...
6  {'Drop Columns Transformer': {'columns': ['cur...
7  {'Drop Columns Transformer': {'columns': ['cur...
8      {'Baseline Classifier': {'strategy': 'mode'}}

```

If we are interested in see more details about the pipeline, we can view a summary description using the `id` from the rankings table:

```
[9]: automl.describe_pipeline(3)
```

```

*****
* Random Forest Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler *
*****

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Drop Columns Transformer
    * columns : ['currency']
2. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
3. DateTime Featurization Component
    * features_to_extract : ['year', 'month', 'day_of_week', 'hour']
    * encode_as_categories : False
    * date_index : None
4. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
5. Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * k_neighbors : 5
6. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6

```

(continues on next page)

(continued from previous page)

```

* n_jobs : -1

Training
=====
Training for binary problems.
Objective to optimize binary classification pipeline thresholds for: <evalml.
↳objectives.standard_metrics.F1 object at 0x7f102ba80af0>
Total training time (including CV): 2.1 seconds

Cross Validation
-----

```

	F1	MCC	Binary	Log	Loss	Binary	Gini	AUC	Precision	Balanced_
↳Accuracy	Binary	Accuracy	Binary	#	Training	#	Validation			
0	0.400	0.476			0.277	0.593	0.797	1.000		
↳	0.625	0.910		133		67				
1	0.364	0.365			0.340	0.318	0.659	0.667		
↳	0.617	0.896		133		67				
2	0.444	0.513			0.253	0.724	0.862	1.000		
↳	0.643	0.924		134		66				
mean	0.403	0.452			0.290	0.545	0.772	0.889		
↳	0.628	0.910		-		-				
std	0.040	0.077			0.045	0.207	0.104	0.192		
↳	0.013	0.014		-		-				
coef of var	0.101	0.170			0.156	0.380	0.134	0.217		
↳	0.021	0.016		-		-				

We can also view the pipeline parameters directly:

```

[10]: pipeline = automl.get_pipeline(3)
print(pipeline.parameters)

{'Drop Columns Transformer': {'columns': ['currency']}, 'Imputer': {'categorical_
↳impute_strategy': 'most_frequent', 'numeric_impute_strategy': 'mean', 'categorical_
↳fill_value': None, 'numeric_fill_value': None}, 'DateTime Featurization Component':
↳{'features_to_extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_
↳categories': False, 'date_index': None}, 'One Hot Encoder': {'top_n': 10, 'features_
↳to_encode': None, 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore
↳', 'handle_missing': 'error'}, 'Oversampler': {'sampling_ratio': 0.25, 'k_neighbors_
↳default': 5, 'n_jobs': -1, 'sampling_ratio_dict': None, 'k_neighbors': 5}, 'Random_
↳Forest Classifier': {'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}

```

We can now select the best pipeline and score it on our holdout data:

```

[11]: pipeline = automl.best_pipeline
pipeline.score(X_holdout, y_holdout, ["f1"])

[11]: OrderedDict([('F1', 0.6666666666666666)])

```

We can also visualize the structure of the components contained by the pipeline:

```

[12]: pipeline.graph()

[12]:

```


TUTORIALS

Below are examples of how to apply EvalML to a variety of problems:

3.1 Building a Fraud Prediction Model with EvalML

In this demo, we will build an optimized fraud prediction model using EvalML. To optimize the pipeline, we will set up an objective function to minimize the percentage of total transaction value lost to fraud. At the end of this demo, we also show you how introducing the right objective during the training results in a much better than using a generic machine learning metric like AUC.

```
[1]: import evalml
      from evalml import AutoMLSearch
      from evalml.objectives import FraudCost
```

3.1.1 Configure “Cost of Fraud”

To optimize the pipelines toward the specific business needs of this model, we can set our own assumptions for the cost of fraud. These parameters are

- `retry_percentage` - what percentage of customers will retry a transaction if it is declined?
- `interchange_fee` - how much of each successful transaction do you collect?
- `fraud_payout_percentage` - the percentage of fraud will you be unable to collect
- `amount_col` - the column in the data the represents the transaction amount

Using these parameters, EvalML determines attempt to build a pipeline that will minimize the financial loss due to fraud.

```
[2]: fraud_objective = FraudCost(retry_percentage=.5,
                                interchange_fee=.02,
                                fraud_payout_percentage=.75,
                                amount_col='amount')
```

3.1.2 Search for best pipeline

In order to validate the results of the pipeline creation and optimization process, we will save some of our data as the holdout set.

```
[3]: X, y = evalml.demos.load_fraud(n_rows=5000)
```

```

                Number of Features
Boolean                      1
Categorical                   6
Numeric                       5

Number of training examples: 5000
Targets
False      86.20%
True       13.80%
Name: fraud, dtype: object
```

EvalML natively supports one-hot encoding. Here we keep 1 out of the 6 categorical columns to decrease computation time.

```
[4]: cols_to_drop = ['datetime', 'expiration_date', 'country', 'region', 'provider']
    for col in cols_to_drop:
        X.ww.pop(col)
```

```
X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
    problem_type='binary', test_size=0.2, random_seed=0)
```

```
X.ww
```

```
[4]:
```

Column	Physical Type	Logical Type	Semantic Tag(s)
card_id	int64	Integer	['numeric']
store_id	int64	Integer	['numeric']
amount	int64	Integer	['numeric']
currency	category	Categorical	['category']
customer_present	bool	Boolean	[]
lat	float64	Double	['numeric']
lng	float64	Double	['numeric']

Because the fraud labels are binary, we will use `AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary')`. When we call `.search()`, the search for the best pipeline will begin.

```
[5]: automl = AutoMLSearch(X_train=X_train, y_train=y_train,
    problem_type='binary',
    objective=fraud_objective,
    additional_objectives=['auc', 'f1', 'precision'],
    allowed_model_families=["random_forest", "linear_model"],
    max_batches=1,
    optimize_thresholds=True,
    verbose=True)
```

```
automl.search()
```

```
Generating pipelines to search over...
3 pipelines ready for search.
```

```
*****
* Beginning pipeline search *
```

(continues on next page)

(continued from previous page)

```

*****

Optimizing for Fraud Cost.
Lower score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 4 pipelines.
Allowed model families: random_forest, linear_model

FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.790

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.539
Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler:
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.289
Logistic Regression Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard_
↪Scaler:
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.539

Search finished after 00:12
Best pipeline: Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler
Best pipeline Fraud Cost: 0.289305

```

View rankings and select pipelines

Once the fitting process is done, we can see all of the pipelines that were searched, ranked by their score on the fraud detection objective we defined.

```

[6]: automl.rankings

[6]:   id  pipeline_name  search_order  \
0    2  Random Forest Classifier w/ Imputer + One Hot ...      2
1    1  Elastic Net Classifier w/ Imputer + One Hot En...      1
2    3  Logistic Regression Classifier w/ Imputer + On...      3
3    0  Mode Baseline Binary Classification Pipeline          0

   mean_cv_score  standard_deviation_cv_score  validation_score  \
0      0.289305          0.434440          0.790953
1      0.538598          0.435507          0.790953
2      0.538598          0.435507          0.790953

```

(continues on next page)

(continued from previous page)

```

3          0.789648          0.001136          0.790953

    percent_better_than_baseline  high_variance_cv  \
0          50.034387          False
1          25.105090          False
2          25.105090          False
3           0.000000          False

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3      {'Baseline Classifier': {'strategy': 'mode'}}

```

To select the best pipeline we can call `automl.best_pipeline`.

```
[7]: best_pipeline = automl.best_pipeline
```

Describe pipelines

We can get more details about any pipeline created during the search process, including how it performed on other objective functions, by calling the `describe_pipeline` method and passing the `id` of the pipeline of interest.

```
[8]: automl.describe_pipeline(automl.rankings.iloc[1]["id"])

*****
* Elastic Net Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard Scaler_
*
*****

Problem Type: binary
Model Family: Linear

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
3. Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * k_neighbors : 5
4. Standard Scaler

```

(continues on next page)

(continued from previous page)

```

5. Elastic Net Classifier
  * penalty : elasticnet
  * C : 1.0
  * l1_ratio : 0.15
  * n_jobs : -1
  * multi_class : auto
  * solver : saga

Training
=====
Training for binary problems.
Objective to optimize binary classification pipeline thresholds for: <evalml.
↳objectives.fraud_cost.FraudCost object at 0x7fa600e64ac0>
Total training time (including CV): 1.9 seconds

Cross Validation
-----

```

	Fraud Cost	AUC	F1	Precision	# Training	# Validation
0	0.791	0.856	0.000	0.000	2,666	1,334
1	0.789	0.796	0.000	0.000	2,667	1,333
2	0.036	0.828	0.635	0.592	2,667	1,333
mean	0.539	0.827	0.212	0.197	-	-
std	0.436	0.030	0.366	0.342	-	-
coef of var	0.809	0.036	1.732	1.732	-	-

3.1.3 Evaluate on holdout data

Finally, since the best pipeline is already trained, we evaluate it on the holdout data.

Now, we can score the pipeline on the holdout data using both our fraud cost objective and the AUC (Area under the ROC Curve) objective.

```

[9]: best_pipeline.score(X_holdout, y_holdout, objectives=["auc", fraud_objective])
[9]: OrderedDict([('AUC', 0.8654628602172233),
                  ('Fraud Cost', 0.026054721586011263)])

```

3.1.4 Why optimize for a problem-specific objective?

To demonstrate the importance of optimizing for the right objective, let's search for another pipeline using AUC, a common machine learning metric. After that, we will score the holdout data using the fraud cost objective to see how the best pipelines compare.

```

[10]: automl_auc = AutoMLSearch(X_train=X_train, y_train=y_train,
                                problem_type='binary',
                                objective='auc',
                                additional_objectives=['f1', 'precision'],
                                max_batches=1,
                                allowed_model_families=["random_forest", "linear_model"],
                                optimize_thresholds=True,
                                verbose=True)

automl_auc.search()

```

```

Generating pipelines to search over...
3 pipelines ready for search.

*****
* Beginning pipeline search *
*****

Optimizing for AUC.
Greater score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 4 pipelines.
Allowed model families: random_forest, linear_model

FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean AUC: 0.500

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.827
Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.847
Logistic Regression Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard_
↪Scaler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.826

Search finished after 00:05
Best pipeline: Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler
Best pipeline AUC: 0.847400

```

Like before, we can look at the rankings of all of the pipelines searched and pick the best pipeline.

```
[11]: automl_auc.rankings
```

```

[11]:   id      pipeline_name  search_order  \
0    2  Random Forest Classifier w/ Imputer + One Hot ...      2
1    1  Elastic Net Classifier w/ Imputer + One Hot En...      1
2    3  Logistic Regression Classifier w/ Imputer + On...      3
3    0      Mode Baseline Binary Classification Pipeline      0

   mean_cv_score  standard_deviation_cv_score  validation_score  \
0      0.847400          0.002923          0.845246
1      0.826778          0.030029          0.856262
2      0.826327          0.030496          0.856172
3      0.500000          0.000000          0.500000

```

(continues on next page)

(continued from previous page)

```

percent_better_than_baseline  high_variance_cv  \
0                          34.740035          False
1                          32.677824          False
2                          32.632734          False
3                          0.000000          False

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3      {'Baseline Classifier': {'strategy': 'mode'}}

```

```
[12]: best_pipeline_auc = automl_auc.best_pipeline
```

```
[13]: # get the fraud score on holdout data
best_pipeline_auc.score(X_holdout, y_holdout, objectives=["auc", fraud_objective])
```

```
[13]: OrderedDict([('AUC', 0.8654628602172233),
                  ('Fraud Cost', 0.026054721586011263)])
```

```
[14]: # fraud score on fraud optimized again
best_pipeline.score(X_holdout, y_holdout, objectives=["auc", fraud_objective])
```

```
[14]: OrderedDict([('AUC', 0.8654628602172233),
                  ('Fraud Cost', 0.026054721586011263)])
```

When we optimize for AUC, we can see that the AUC score from this pipeline performs better compared to the AUC score from the pipeline optimized for fraud cost; however, the losses due to fraud are a much larger percentage of the total transaction amount when optimized for AUC and much smaller when optimized for fraud cost. As a result, we lose a noticeable percentage of the total transaction amount by not optimizing for fraud cost specifically.

Optimizing for AUC does not take into account the user-specified `retry_percentage`, `interchange_fee`, `fraud_payout_percentage` values, which could explain the decrease in fraud performance. Thus, the best pipelines may produce the highest AUC but may not actually reduce the amount loss due to your specific type fraud.

This example highlights how performance in the real world can diverge greatly from machine learning metrics.

3.2 Building a Lead Scoring Model with EvalML

In this demo, we will build an optimized lead scoring model using EvalML. To optimize the pipeline, we will set up an objective function to maximize the revenue generated with true positives while taking into account the cost of false positives. At the end of this demo, we also show you how introducing the right objective during the training is significantly better than using a generic machine learning metric like AUC.

```
[1]: import evalml
from evalml import AutoMLSearch
from evalml.objectives import LeadScoring
```

3.2.1 Configure LeadScoring

To optimize the pipelines toward the specific business needs of this model, you can set your own assumptions for how much value is gained through true positives and the cost associated with false positives. These parameters are

- `true_positive` - dollar amount to be gained with a successful lead
- `false_positive` - dollar amount to be lost with an unsuccessful lead

Using these parameters, EvalML builds a pipeline that will maximize the amount of revenue per lead generated.

```
[2]: lead_scoring_objective = LeadScoring(
      true_positives=100,
      false_positives=-5
    )
```

3.2.2 Dataset

We will be utilizing a dataset detailing a customer's job, country, state, zip, online action, the dollar amount of that action and whether they were a successful lead.

```
[3]: from urllib.request import urlopen
      import pandas as pd
      import woodwork as ww
      customers_data = urlopen('https://featurelabs-static.s3.amazonaws.com/lead_scoring_ml_
      ↪apps/customers.csv')
      interactions_data = urlopen('https://featurelabs-static.s3.amazonaws.com/lead_scoring_
      ↪ml_apps/interactions.csv')
      leads_data = urlopen('https://featurelabs-static.s3.amazonaws.com/lead_scoring_ml_
      ↪apps/previous_leads.csv')
      customers = pd.read_csv(customers_data)
      interactions = pd.read_csv(interactions_data)
      leads = pd.read_csv(leads_data)
```

```
X = customers.merge(interactions, on='customer_id').merge(leads, on='customer_id')
y = X['label']
X = X.drop(['customer_id', 'date_registered', 'birthday', 'phone', 'email',
           'owner', 'company', 'id', 'time_x',
           'session', 'referrer', 'time_y', 'label', 'country'], axis=1)
display(X.head())
```

	job	state	zip	action	amount
0	Engineer, mining	NY	60091.0	page_view	NaN
1	Psychologist, forensic	CA	NaN	purchase	135.23
2	Psychologist, forensic	CA	NaN	page_view	NaN
3	Air cabin crew	NaN	60091.0	download	NaN
4	Air cabin crew	NaN	60091.0	page_view	NaN

We will convert our data into Woodwork data structures. Doing so enables us to have more control over the types passed to and inferred by AutoML.

```
[4]: X.ww.init(semantic_tags={'job': 'category'}, logical_types={'job': 'Categorical'})
      y = ww.init_series(y)
      X.ww
```

```
[4]: Physical Type Logical Type Semantic Tag(s)
Column
job          category  Categorical      ['category']
```

(continues on next page)

(continued from previous page)

state	category	Categorical	['category']
zip	float64	Double	['numeric']
action	category	Categorical	['category']
amount	float64	Double	['numeric']

3.2.3 Search for the best pipeline

In order to validate the results of the pipeline creation and optimization process, we will save some of our data as a holdout set.

EvalML natively supports one-hot encoding and imputation so the above NaN and categorical values will be taken care of.

```
[5]: X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
    ↪problem_type='binary', test_size=0.2, random_seed=0)
```

```
X.ww
```

```
[5]:      Physical Type Logical Type Semantic Tag(s)
Column
job      category  Categorical  ['category']
state    category  Categorical  ['category']
zip      float64    Double       ['numeric']
action   category    Categorical  ['category']
amount   float64      Double       ['numeric']
```

Because the lead scoring labels are binary, we will use set the problem type to “binary”. When we call `.search()`, the search for the best pipeline will begin.

```
[6]: automl = AutoMLSearch(X_train=X_train, y_train=y_train,
    ↪problem_type='binary',
    ↪objective=lead_scoring_objective,
    ↪additional_objectives=['auc'],
    ↪allowed_model_families=["catboost", "random_forest", "linear_
    ↪model"],
    ↪max_batches=1,
    ↪verbose=True)
```

```
automl.search()
```

```
Generating pipelines to search over...
```

```
4 pipelines ready for search.
```

```
*****
* Beginning pipeline search *
*****
```

```
Optimizing for Lead Scoring.
Greater score is better.
```

```
Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 5 pipelines.
Allowed model families: linear_model, catboost, random_forest
```

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.000

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.028
Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler:
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.031
Logistic Regression Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard_
↪Scaler:
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.028
CatBoost Classifier w/ Imputer + Oversampler:
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.358

Search finished after 00:11
Best pipeline: CatBoost Classifier w/ Imputer + Oversampler
Best pipeline Lead Scoring: 0.358194
```

View rankings and select pipeline

Once the fitting process is done, we can see all of the pipelines that were searched, ranked by their score on the lead scoring objective we defined.

```
[7]: automl.rankings
```

```
[7]:   id                pipeline_name  search_order  \
0    4      CatBoost Classifier w/ Imputer + Oversampler      4
1    2  Random Forest Classifier w/ Imputer + One Hot ...      2
2    1  Elastic Net Classifier w/ Imputer + One Hot En...      1
3    3  Logistic Regression Classifier w/ Imputer + On...      3
4    0      Mode Baseline Binary Classification Pipeline      0

   mean_cv_score  standard_deviation_cv_score  validation_score  \
0      0.358194          0.328811          -0.016129
1      0.031193          0.041485          -0.016129
2      0.027979          0.065303          -0.012903
3      0.027979          0.065303          -0.012903
4      0.000000          0.000000          0.000000

   percent_better_than_baseline  high_variance_cv  \
0                inf                False
1                inf                False
2                inf                False
```

(continues on next page)

(continued from previous page)

```

3             inf             False
4             0.0            False

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3  {'Imputer': {'categorical_impute_strategy': 'm...
4      {'Baseline Classifier': {'strategy': 'mode'}}

```

To select the best pipeline we can call `automl.best_pipeline`.

```
[8]: best_pipeline = automl.best_pipeline
```

Describe pipeline

You can get more details about any pipeline, including how it performed on other objective functions by calling `.describe_pipeline()` and specifying the `id` of the pipeline.

```
[9]: automl.describe_pipeline(automl.rankings.iloc[0]["id"])
```

```

*****
* CatBoost Classifier w/ Imputer + Oversampler *
*****

Problem Type: binary
Model Family: CatBoost

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * categorical_features : [0, 1, 3]
    * k_neighbors : 5
3. CatBoost Classifier
    * n_estimators : 10
    * eta : 0.03
    * max_depth : 6
    * bootstrap_type : None
    * silent : True
    * allow_writing_files : False
    * n_jobs : -1

Training
=====
Training for binary problems.

```

(continues on next page)

(continued from previous page)

```
Objective to optimize binary classification pipeline thresholds for: <evalml.
↳objectives.lead_scoring.LeadScoring object at 0x7f63a3231a60>
Total training time (including CV): 1.2 seconds
```

```
Cross Validation
```

```
-----
Lead Scoring    AUC # Training # Validation
0               -0.016 0.869      3,099      1,550
1               0.490 0.887      3,099      1,550
2               0.600 0.889      3,100      1,549
mean            0.358 0.882          -          -
std             0.329 0.011          -          -
coef of var     0.918 0.012          -          -
```

3.2.4 Evaluate on hold out

Finally, since the best pipeline was trained on all of the training data, we evaluate it on the holdout dataset.

```
[10]: best_pipeline_score = best_pipeline.score(X_holdout, y_holdout, objectives=["auc",
↳lead_scoring_objective])
best_pipeline_score

[10]: OrderedDict([('AUC', 0.8585599879117558),
('Lead Scoring', 0.8469475494411006)])
```

3.2.5 Why optimize for a problem-specific objective?

To demonstrate the importance of optimizing for the right objective, let's search for another pipeline using AUC, a common machine learning metric. After that, we will score the holdout data using the lead scoring objective to see how the best pipelines compare.

```
[11]: automl_auc = evalml.AutoMLSearch(X_train=X_train, y_train=y_train,
problem_type='binary',
objective='auc',
additional_objectives=[lead_scoring_objective],
allowed_model_families=["catboost", "random_forest",
↳"linear_model"],
max_batches=1,
verbose=True)

automl_auc.search()

Generating pipelines to search over...
4 pipelines ready for search.

*****
* Beginning pipeline search *
*****

Optimizing for AUC.
Greater score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 5 pipelines.
```

(continues on next page)

(continued from previous page)

```
Allowed model families: linear_model, catboost, random_forest
```

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...}
```

```
Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
```

```
Mode Baseline Binary Classification Pipeline:
```

```
Starting cross validation
```

```
Finished cross validation - mean AUC: 0.500
```

```
*****
```

```
* Evaluating Batch Number 1 *
```

```
*****
```

```
Elastic Net Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard Scaler:
```

```
Starting cross validation
```

```
Finished cross validation - mean AUC: 0.672
```

```
Random Forest Classifier w/ Imputer + One Hot Encoder + Oversampler:
```

```
Starting cross validation
```

```
Finished cross validation - mean AUC: 0.662
```

```
Logistic Regression Classifier w/ Imputer + One Hot Encoder + Oversampler + Standard_
↪Scaler:
```

```
Starting cross validation
```

```
Finished cross validation - mean AUC: 0.672
```

```
CatBoost Classifier w/ Imputer + Oversampler:
```

```
Starting cross validation
```

```
Finished cross validation - mean AUC: 0.882
```

```
Search finished after 00:07
```

```
Best pipeline: CatBoost Classifier w/ Imputer + Oversampler
```

```
Best pipeline AUC: 0.881709
```

```
[12]: automl_auc.rankings
```

```
[12]:
```

	id	pipeline_name	search_order	\
0	4	CatBoost Classifier w/ Imputer + Oversampler	4	
1	1	Elastic Net Classifier w/ Imputer + One Hot En...	1	
2	3	Logistic Regression Classifier w/ Imputer + On...	3	
3	2	Random Forest Classifier w/ Imputer + One Hot ...	2	
4	0	Mode Baseline Binary Classification Pipeline	0	

	mean_cv_score	standard_deviation_cv_score	validation_score	\
0	0.881709	0.010861	0.869201	
1	0.671885	0.050182	0.727798	
2	0.671517	0.048732	0.725816	
3	0.662322	0.050673	0.718588	
4	0.500000	0.000000	0.500000	

	percent_better_than_baseline	high_variance_cv	\
0	38.170947	False	
1	17.188511	False	
2	17.151664	False	
3	16.232241	False	
4	0.000000	False	

(continues on next page)

(continued from previous page)

```

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3  {'Imputer': {'categorical_impute_strategy': 'm...
4      {'Baseline Classifier': {'strategy': 'mode'}}

```

Like before, we can look at the rankings and pick the best pipeline.

```
[13]: best_pipeline_auc = automl_auc.best_pipeline
```

```
[14]: # get the auc and lead scoring score on holdout data
best_pipeline_auc_score = best_pipeline_auc.score(X_holdout, y_holdout, objectives=[
    ↪ "auc", lead_scoring_objective])
best_pipeline_auc_score
```

```
[14]: OrderedDict([('AUC', 0.8585599879117558),
                  ('Lead Scoring', 0.08168529664660361)])
```

```
[15]: assert best_pipeline_score['Lead Scoring'] > best_pipeline_auc_score['Lead Scoring']
assert best_pipeline_auc_score['Lead Scoring'] >= 0
```

When we optimize for AUC, we can see that the AUC score from this pipeline is similar to the AUC score from the pipeline optimized for lead scoring. However, the revenue per lead is much smaller per lead when optimized for AUC and was much larger when optimized for lead scoring. As a result, we would have a huge gain on the amount of revenue if we optimized for lead scoring.

This happens because optimizing for AUC does not take into account the user-specified `true_positive` (dollar amount to be gained with a successful lead) and `false_positive` (dollar amount to be lost with an unsuccessful lead) values. Thus, the best pipelines may produce the highest AUC but may not actually generate the most revenue through lead scoring.

This example highlights how performance in the real world can diverge greatly from machine learning metrics.

3.3 Using the Cost-Benefit Matrix Objective

The Cost-Benefit Matrix (`CostBenefitMatrix`) objective is an objective that assigns costs to each of the quadrants of a confusion matrix to quantify the cost of being correct or incorrect.

3.3.1 Confusion Matrix

Confusion matrices are tables that summarize the number of correct and incorrectly-classified predictions, broken down by each class. They allow us to quickly understand the performance of a classification model and where the model gets “confused” when it is making predictions. For the binary classification problem, there are four possible combinations of prediction and actual target values possible:

- true positives (correct positive assignments)
- true negatives (correct negative assignments)
- false positives (incorrect positive assignments)
- false negatives (incorrect negative assignments)

An example of how to calculate a confusion matrix can be found [here](#).

3.3.2 Cost-Benefit Matrix

Although the confusion matrix is an incredibly useful visual for understanding our model, each prediction that is correctly or incorrectly classified is treated equally. For example, for detecting breast cancer, the confusion matrix does not take into consideration that it could be much more costly to incorrectly classify a malignant tumor as benign than it is to incorrectly classify a benign tumor as malignant. This is where the cost-benefit matrix shines: it uses the cost of each of the four possible outcomes to weigh each outcome differently. By scoring using the cost-benefit matrix, we can measure the score of the model by a concrete unit that is more closely related to the goal of the model. In the below example, we will show how the cost-benefit matrix objective can be used, and how it can give us better real-world impact when compared to using other standard machine learning objectives.

3.3.3 Customer Churn Example

Data

In this example, we will be using a customer churn data set taken from [Kaggle](#).

This dataset includes records of over 7000 customers, and includes customer account information, demographic information, services they signed up for, and whether or not the customer “churned” or left within the last month.

The target we want to predict is whether the customer churned (“Yes”) or did not churn (“No”). In the dataset, approximately 73.5% of customers did not churn, and 26.5% did. We will refer to the customers who churned as the “positive” class and the customers who did not churn as the “negative” class.

```
[1]: from evalml.demos.churn import load_churn
      from evalml.preprocessing import split_data

X, y = load_churn()
X.ww.set_types({'PaymentMethod': 'Categorical', 'Contract': 'Categorical'}) # Update_
    ↪ data types Woodwork did not correctly infer
X_train, X_holdout, y_train, y_holdout = split_data(X, y, problem_type='binary', test_
    ↪ size=0.3, random_seed=0)
```

	Number of Features
Categorical	16
Numeric	3

```
Number of training examples: 7043
Targets
No      73.46%
Yes     26.54%
Name: Churn, dtype: object
```

In this example, let’s say that correctly identifying customers who will churn (true positive case) will give us a net profit of \$400, because it allows us to intervene, incentivize the customer to stay, and sign a new contract. Incorrectly classifying customers who were not going to churn as customers who will churn (false positive case) will cost \$100 to represent the marketing and effort used to try to retain the user. Not identifying customers who will churn (false negative case) will cost us \$200 to represent the lost in revenue from losing a customer. Finally, correctly identifying customers who will not churn (true negative case) will not cost us anything (\$0), as nothing needs to be done for that customer.

We can represent these values in our `CostBenefitMatrix` objective, where a negative value represents a cost and a positive value represents a profit—note that this means that the greater the score, the more profit we will make.

```
[2]: from evalml.objectives import CostBenefitMatrix
      cost_benefit_matrix = CostBenefitMatrix(true_positive=400,
```

(continues on next page)

(continued from previous page)

```
true_negative=0,
false_positive=-100,
false_negative=-200)
```

AutoML Search with Log Loss

First, let us run AutoML search to train pipelines using the default objective for binary classification (log loss).

```
[3]: from evalml import AutoMLSearch
automl = AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary',
    ↳objective='log loss binary',
    ↳max_iterations=5, verbose=True)
automl.search()

ll_pipeline = automl.best_pipeline
ll_pipeline.score(X_holdout, y_holdout, ['log loss binary'])

Generating pipelines to search over...
Removing columns ['TotalCharges'] because they are of 'Unknown' type
8 pipelines ready for search.

*****
* Beginning pipeline search *
*****

Optimizing for Log Loss Binary.
Lower score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 5 pipelines.
Allowed model families: extra_trees, decision_tree, random_forest, linear_model,
    ↳xgboost, lightgbm, catboost

FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type': ...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 9.164

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder +
    ↳Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.423
Decision Tree Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.685
Random Forest Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
```

(continues on next page)

(continued from previous page)

```

    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.426
LightGBM Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.471

Search finished after 00:10
Best pipeline: Elastic Net Classifier w/ Drop Columns Transformer + Imputer + One Hot
↳Encoder + Standard Scaler
Best pipeline Log Loss Binary: 0.422746
[3]: OrderedDict([('Log Loss Binary', 0.41990933380906487)])

```

When we train our pipelines using log loss as our primary objective, we try to find pipelines that minimize log loss. However, our ultimate goal in training models is to find a model that gives us the most profit, so let's score our pipeline on the cost benefit matrix (using the costs outlined above) to determine the profit we would earn from the predictions made by this model:

```

[4]: ll_pipeline_score = ll_pipeline.score(X_holdout, y_holdout, [cost_benefit_matrix])
print (ll_pipeline_score)

OrderedDict([('Cost Benefit Matrix', 54.47231424514908)])

[5]: # Calculate total profit across all customers using pipeline optimized for Log Loss
total_profit_ll = ll_pipeline_score['Cost Benefit Matrix'] * len(X)
print (total_profit_ll)

383648.50922858494

```

AutoML Search with Cost-Benefit Matrix

Let's try rerunning our AutoML search, but this time using the cost-benefit matrix as our primary objective to optimize.

```

[6]: automl = AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary',
↳objective=cost_benefit_matrix,
                                max_iterations=5, verbose=True)
automl.search()

cbm_pipeline = automl.best_pipeline

Generating pipelines to search over...
Removing columns ['TotalCharges'] because they are of 'Unknown' type
8 pipelines ready for search.

*****
* Beginning pipeline search *
*****

Optimizing for Cost Benefit Matrix.
Greater score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 5 pipelines.
Allowed model families: extra_trees, decision_tree, random_forest, linear_model,
↳xgboost, lightgbm, catboost

```

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Cost Benefit Matrix: -53.063

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder +
↳Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Cost Benefit Matrix: 59.554
Decision Tree Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Cost Benefit Matrix: 51.745
Random Forest Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Cost Benefit Matrix: 59.413
LightGBM Classifier w/ Drop Columns Transformer + Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Cost Benefit Matrix: 50.692

Search finished after 00:14
Best pipeline: Elastic Net Classifier w/ Drop Columns Transformer + Imputer + One Hot
↳Encoder + Standard Scaler
Best pipeline Cost Benefit Matrix: 59.554148
```

Now, if we calculate the cost-benefit matrix score on our best pipeline, we see that with this pipeline optimized for our cost-benefit matrix objective, we are able to generate more profit per customer. Across our 7043 customers, we generate much more profit using this best pipeline! Custom objectives like `CostBenefitMatrix` are just one example of how using EvalML can help find pipelines that can perform better on real-world problems, rather than on arbitrary standard statistical metrics.

```
[7]: cbm_pipeline_score = cbm_pipeline.score(X_holdout, y_holdout, [cost_benefit_matrix])
print (cbm_pipeline_score)
```

```
OrderedDict([('Cost Benefit Matrix', 62.23379081874113)])
```

```
[8]: # Calculate total profit across all customers using pipeline optimized for
↳CostBenefitMatrix
total_profit_cbm = cbm_pipeline_score['Cost Benefit Matrix'] * len(X)
print (total_profit_cbm)
```

```
438312.5887363938
```

```
[9]: # Calculate difference in profit made using both pipelines
profit_diff = total_profit_cbm - total_profit_ll
print (profit_diff)
```

```
54664.07950780884
```

Finally, we can graph the confusion matrices for both pipelines to better understand why the pipeline trained using the cost-benefit matrix is able to correctly classify more samples than the pipeline trained with log loss: we were able

to correctly predict more cases where the customer would have churned (true positive), allowing us to intervene and prevent those customers from leaving.

```
[10]: from evalml.model_understanding.graphs import graph_confusion_matrix

# pipeline trained with log loss
y_pred = ll_pipeline.predict(X_holdout)
graph_confusion_matrix(y_holdout, y_pred)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

```
[11]: # pipeline trained with cost-benefit matrix
y_pred = cbm_pipeline.predict(X_holdout)
graph_confusion_matrix(y_holdout, y_pred)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

3.4 Using Text Data with EvalML

In this demo, we will show you how to use EvalML to build models which use text data.

```
[1]: import evalml
from evalml import AutoMLSearch
```

3.4.1 Dataset

We will be utilizing a dataset of SMS text messages, some of which are categorized as spam, and others which are not (“ham”). This dataset is originally from [Kaggle](#), but modified to produce a slightly more even distribution of spam to ham.

```
[2]: from urllib.request import urlopen
import pandas as pd

input_data = urlopen('https://featurelabs-static.s3.amazonaws.com/spam_text_messages_
↳modified.csv')
data = pd.read_csv(input_data)[:750]

X = data.drop(['Category'], axis=1)
y = data['Category']

display(X.head())
```

```

                                Message
0  Free entry in 2 a wkly comp to win FA Cup fina...
1  FreeMsg Hey there darling it's been 3 week's n...
2  WINNER!! As a valued network customer you have...
3  Had your mobile 11 months or more? U R entitle...
4  SIX chances to win CASH! From 100 to 20,000 po...
```

The ham vs spam distribution of the data is 3:1, so any machine learning model must get above 75% accuracy in order to perform better than a trivial baseline model which simply classifies everything as ham.

```
[3]: y.value_counts(normalize=True)

[3]: spam    0.593333
     ham    0.406667
     Name: Category, dtype: float64
```

In order to properly utilize Woodwork's 'Natural Language' typing, we need to pass this argument in during initialization. Otherwise, this will be treated as an 'Unknown' type and dropped in the search.

```
[4]: X.ww.init(logical_types={"Message": "NaturalLanguage"})
```

3.4.2 Search for best pipeline

In order to validate the results of the pipeline creation and optimization process, we will save some of our data as a holdout set.

```
[5]: X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
    ↪problem_type='binary', test_size=0.2, random_seed=0)
```

EvalML uses Woodwork to automatically detect which columns are text columns, so you can run search normally, as you would if there was no text data. We can print out the logical type of the Message column and assert that it is indeed inferred as a natural language column.

```
[6]: X_train.ww

[6]:      Physical Type      Logical Type Semantic Tag(s)
     Column
     Message      string NaturalLanguage      []
```

Because the spam/ham labels are binary, we will use `AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary')`. When we call `.search()`, the search for the best pipeline will begin.

```
[7]: automl = AutoMLSearch(X_train=X_train, y_train=y_train,
    problem_type='binary',
    max_batches=1,
    optimize_thresholds=True,
    verbose=True)
```

```
automl.search()
```

```
Generating pipelines to search over...
8 pipelines ready for search.
```

```
*****
* Beginning pipeline search *
*****
```

```
Optimizing for Log Loss Binary.
Lower score is better.
```

```
Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 9 pipelines.
```

(continues on next page)

(continued from previous page)

```

Allowed model families: catboost, extra_trees, decision_tree, lightgbm, xgboost,
↳random_forest, linear_model

FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 14.046

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Text Featurization Component + Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
Decision Tree Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 3.386
  High coefficient of variation (cv >= 0.5) within cross validation scores.
  Decision Tree Classifier w/ Text Featurization Component + Imputer may not
↳perform as estimated on unseen data.
Random Forest Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.221
LightGBM Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
Logistic Regression Classifier w/ Text Featurization Component + Imputer + Standard
↳Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
XGBoost Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.266
Extra Trees Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
CatBoost Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.589

Search finished after 00:27
Best pipeline: Random Forest Classifier w/ Text Featurization Component + Imputer
Best pipeline Log Loss Binary: 0.221422

```

View rankings and select pipeline

Once the fitting process is done, we can see all of the pipelines that were searched.

```
[8]: automl.rankings
```

```
[8]:   id      pipeline_name  search_order  \
0    3  Random Forest Classifier w/ Text Featurization...    3
1    6  XGBoost Classifier w/ Text Featurization Compo...    6
2    4  LightGBM Classifier w/ Text Featurization Comp...    4
3    7  Extra Trees Classifier w/ Text Featurization C...    7
4    5  Logistic Regression Classifier w/ Text Featuri...    5
5    1  Elastic Net Classifier w/ Text Featurization C...    1
6    8  CatBoost Classifier w/ Text Featurization Comp...    8
7    2  Decision Tree Classifier w/ Text Featurization...    2
8    0      Mode Baseline Binary Classification Pipeline    0

   mean_cv_score  standard_deviation_cv_score  validation_score  \
0      0.221422                0.040958          0.221587
1      0.266164                0.106501          0.242896
2      0.291768                0.114862          0.291521
3      0.292373                0.029893          0.325764
4      0.350340                0.074833          0.349271
5      0.350471                0.074886          0.349437
6      0.588944                0.004016          0.592259
7      3.385551                0.672118          3.708759
8     14.045769                0.099705         13.988204

   percent_better_than_baseline  high_variance_cv  \
0                        98.423568          False
1                        98.105025          False
2                        97.922737          False
3                        97.918427          False
4                        97.505728          False
5                        97.504795          False
6                        95.806967          False
7                        75.896294           True
8                        0.000000          False

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3  {'Imputer': {'categorical_impute_strategy': 'm...
4  {'Imputer': {'categorical_impute_strategy': 'm...
5  {'Imputer': {'categorical_impute_strategy': 'm...
6  {'Imputer': {'categorical_impute_strategy': 'm...
7  {'Imputer': {'categorical_impute_strategy': 'm...
8      {'Baseline Classifier': {'strategy': 'mode'}}
```

To select the best pipeline we can call `automl.best_pipeline`.

```
[9]: best_pipeline = automl.best_pipeline
```

Describe pipeline

You can get more details about any pipeline, including how it performed on other objective functions.

```
[10]: automl.describe_pipeline(automl.rankings.iloc[0]["id"])

*****
* Random Forest Classifier w/ Text Featurization Component + Imputer *
*****

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Text Featurization Component
2. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
3. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6
    * n_jobs : -1

Training
=====
Training for binary problems.
Total training time (including CV): 3.4 seconds

Cross Validation
-----
          Log Loss Binary  MCC Binary  Gini  AUC  Precision  F1  Balanced_
↳Accuracy Binary  Accuracy Binary # Training # Validation
0          0.222          0.817 0.950 0.975          0.862 0.893
↳          0.913          0.910          400          200
1          0.180          0.875 0.970 0.985          0.937 0.925
↳          0.936          0.940          400          200
2          0.262          0.783 0.925 0.963          0.918 0.865
↳          0.883          0.895          400          200
mean          0.221          0.825 0.948 0.974          0.906 0.894
↳          0.910          0.915          -          -
std          0.041          0.047 0.023 0.011          0.039 0.030
↳          0.026          0.023          -          -
coef of var          0.185          0.057 0.024 0.012          0.043 0.034
↳          0.029          0.025          -          -
```

```
[11]: best_pipeline.graph()
```

```
[11]:
```

Notice above that there is a `Text Featurization Component` as the first step in the pipeline. `AutoMLSearch` uses the `woodwork` accessor to recognize that `'Message'` is a text column, and converts this text into numerical values that can be handled by the estimator.

3.4.3 Evaluate on holdout

Now, we can score the pipeline on the holdout data using the core objectives for binary classification problems.

```
[12]: scores = best_pipeline.score(X_holdout, y_holdout, objectives=evalml.objectives.get_
      ↪core_objectives('binary'))
      print(f'Accuracy Binary: {scores["Accuracy Binary"]}')

Accuracy Binary: 0.96
```

As you can see, this model performs relatively well on this dataset, even on unseen data.

3.4.4 What does the Text Featurization Component do?

Machine learning models cannot handle non-numeric data. Any text must be broken down into numeric features that provide useful information about that text. The Text Featurization component first normalizes your text by removing any punctuation and other non-alphanumeric characters and converting any capital letters to lowercase. From there, it passes the text into `featuretools`' `nlp_primitives` dfs search, resulting in several informative features that replace the original column in your dataset: Diversity Score, Mean Characters per Word, Polarity Score, and LSA (Latent Semantic Analysis).

Diversity Score is the ratio of unique words to total words.

Mean Characters per Word is the average number of letters in each word.

Polarity Score is a prediction of how “polarized” the text is, on a scale from -1 (extremely negative) to 1 (extremely positive).

Latent Semantic Analysis is an abstract representation of how important each word is with respect to the entire text, reduced down into two values per text. While the other text features are each a single column, this feature adds two columns to your data, `LSA(column_name) [0]` and `LSA(column_name) [1]`.

Let's see what this looks like with our spam/ham example.

```
[13]: best_pipeline.input_feature_names

[13]: {'Text Featurization Component': ['Message'],
      'Imputer': ['DIVERSITY_SCORE(Message)',
                  'MEAN_CHARACTERS_PER_WORD(Message)',
                  'POLARITY_SCORE(Message)',
                  'LSA(Message) [0]',
                  'LSA(Message) [1]'],
      'Random Forest Classifier': ['DIVERSITY_SCORE(Message)',
                                   'MEAN_CHARACTERS_PER_WORD(Message)',
                                   'POLARITY_SCORE(Message)',
                                   'LSA(Message) [0]',
                                   'LSA(Message) [1]']}
```

Here, the Text Featurization component takes in a single “Message” column, but then the next component in the pipeline, the Imputer, receives five columns of input. These five columns are the result of featurizing the text-type “Message” column. Most importantly, these featurized columns are what ends up passed in to the estimator.

If the dataset had any non-text columns, those would be left alone by this process. If the dataset had more than one text column, each would be broken into these five feature columns independently.

The features, more directly

Rather than just checking the new column names, let's examine the output of this component directly. We can see this by running the component on its own.

```
[14]: text_featurizer = evalml.pipelines.components.TextFeaturizer()
      X_featurized = text_featurizer.fit_transform(X_train)
```

Now we can compare the input data to the output from the text featurizer:

```
[15]: X_train.head()
```

```
[15]:      Message
296  Sunshine Hols. To claim ur med holiday send a ...
652      Yup ü not comin :-(
526  Hello hun how ru? Its here by the way. Im good...
571  I tagged MY friends that you seemed to count a...
472      What happened to our yo date?
```

```
[16]: X_featurized.head()
```

```
[16]:      DIVERSITY_SCORE (Message)  MEAN_CHARACTERS_PER_WORD (Message)  \
296                        1.0                        4.344828
652                        1.0                        3.000000
526                        1.0                        3.363636
571                        0.8                        4.083333
472                        1.0                        3.833333

      POLARITY_SCORE (Message)  LSA (Message) [0]  LSA (Message) [1]
296                        0.003      0.150556      -0.072443
652                        0.000      0.017340      -0.005411
526                        0.162      0.169954      0.022670
571                        0.681      0.144713      0.036799
472                        0.000      0.109373      -0.042754
```

These numeric values now represent important information about the original text that the estimator at the end of the pipeline can successfully use to make predictions.

3.4.5 Why encode text this way?

To demonstrate the importance of text-specific modeling, let's train a model with the same dataset, without letting AutoMLSearch detect the text column. We can change this by explicitly setting the data type of the 'Message' column in Woodwork to Categorical using the utility method `infer_feature_types`.

```
[17]: from evalml.utils import infer_feature_types
      X = infer_feature_types(X, {'Message': 'Categorical'})
      X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
      ↪problem_type='binary', test_size=0.2, random_seed=0)
```

```
[18]: automl_no_text = AutoMLSearch(X_train=X_train, y_train=y_train,
      problem_type='binary',
      max_batches=1,
      optimize_thresholds=True,
      verbose=True)

      automl_no_text.search()
```

```
Generating pipelines to search over...
8 pipelines ready for search.
```

```
*****
* Beginning pipeline search *
*****
```

```
Optimizing for Log Loss Binary.
Lower score is better.
```

```
Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 9 pipelines.
Allowed model families: catboost, extra_trees, decision_tree, lightgbm, xgboost,
↳random_forest, linear_model
```

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...}]
```

```
Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 14.046
```

```
*****
* Evaluating Batch Number 1 *
*****
```

```
Elastic Net Classifier w/ Text Featurization Component + Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
Decision Tree Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 3.386
  High coefficient of variation (cv >= 0.5) within cross validation scores.
  Decision Tree Classifier w/ Text Featurization Component + Imputer may not
↳perform as estimated on unseen data.
Random Forest Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.221
LightGBM Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
Logistic Regression Classifier w/ Text Featurization Component + Imputer + Standard
↳Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
XGBoost Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.266
Extra Trees Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
CatBoost Classifier w/ Text Featurization Component + Imputer:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.589
```

(continues on next page)

(continued from previous page)

```

Search finished after 00:24
Best pipeline: Random Forest Classifier w/ Text Featurization Component + Imputer
Best pipeline Log Loss Binary: 0.221422

```

Like before, we can look at the rankings and pick the best pipeline.

```
[19]: automl_no_text.rankings
```

```

[19]:   id      pipeline_name  search_order  \
0    3  Random Forest Classifier w/ Text Featurization...      3
1    6  XGBoost Classifier w/ Text Featurization Compo...      6
2    4  LightGBM Classifier w/ Text Featurization Comp...      4
3    7  Extra Trees Classifier w/ Text Featurization C...      7
4    5  Logistic Regression Classifier w/ Text Featuri...      5
5    1  Elastic Net Classifier w/ Text Featurization C...      1
6    8  CatBoost Classifier w/ Text Featurization Comp...      8
7    2  Decision Tree Classifier w/ Text Featurization...      2
8    0      Mode Baseline Binary Classification Pipeline      0

   mean_cv_score  standard_deviation_cv_score  validation_score  \
0         0.221422             0.040958             0.221587
1         0.266164             0.106501             0.242896
2         0.291768             0.114862             0.291521
3         0.292373             0.029893             0.325764
4         0.350340             0.074833             0.349271
5         0.350471             0.074886             0.349437
6         0.588944             0.004016             0.592259
7         3.385551             0.672118             3.708759
8        14.045769             0.099705            13.988204

   percent_better_than_baseline  high_variance_cv  \
0                98.423568             False
1                98.105025             False
2                97.922737             False
3                97.918427             False
4                97.505728             False
5                97.504795             False
6                95.806967             False
7                75.896294              True
8                 0.000000             False

                                parameters
0  {'Imputer': {'categorical_impute_strategy': 'm...
1  {'Imputer': {'categorical_impute_strategy': 'm...
2  {'Imputer': {'categorical_impute_strategy': 'm...
3  {'Imputer': {'categorical_impute_strategy': 'm...
4  {'Imputer': {'categorical_impute_strategy': 'm...
5  {'Imputer': {'categorical_impute_strategy': 'm...
6  {'Imputer': {'categorical_impute_strategy': 'm...
7  {'Imputer': {'categorical_impute_strategy': 'm...
8      {'Baseline Classifier': {'strategy': 'mode'}}

```

```
[20]: best_pipeline_no_text = automl_no_text.best_pipeline
```

Here, changing the data type of the text column removed the Text Featurization Component from the pipeline.

```
[21]: best_pipeline_no_text.graph()
```

```
[21]:
```

```
[22]: automl_no_text.describe_pipeline(automl_no_text.rankings.iloc[0]["id"])

*****
* Random Forest Classifier w/ Text Featurization Component + Imputer *
*****

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Text Featurization Component
2. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
3. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6
    * n_jobs : -1

Training
=====
Training for binary problems.
Total training time (including CV): 3.4 seconds

Cross Validation
-----
              Log Loss Binary  MCC Binary  Gini  AUC  Precision  F1  Balanced_
↳Accuracy Binary  Accuracy Binary # Training # Validation
0              0.222          0.817 0.950 0.975          0.862 0.893
↳          0.913          0.910          400          200
1              0.180          0.875 0.970 0.985          0.937 0.925
↳          0.936          0.940          400          200
2              0.262          0.783 0.925 0.963          0.918 0.865
↳          0.883          0.895          400          200
mean           0.221          0.825 0.948 0.974          0.906 0.894
↳          0.910          0.915          -          -
std            0.041          0.047 0.023 0.011          0.039 0.030
↳          0.026          0.023          -          -
coef of var     0.185          0.057 0.024 0.012          0.043 0.034
↳          0.029          0.025          -          -
```

```
[23]: # get standard performance metrics on holdout data
scores = best_pipeline_no_text.score(X_holdout, y_holdout, objectives=evalml.
↳objectives.get_core_objectives('binary'))
print(f'Accuracy Binary: {scores["Accuracy Binary"]}')

Accuracy Binary: 0.96
```

Without the Text Featurization Component, the 'Message' column was treated as a categorical column, and therefore the conversion of this text to numerical features happened in the One Hot Encoder. The best pipeline encoded the top 10 most frequent “categories” of these texts, meaning 10 text messages were one-hot encoded and

all the others were dropped. Clearly, this removed almost all of the information from the dataset, as we can see the `best_pipeline_no_text` performs very similarly to randomly guessing “ham” in every case.

These guides include in-depth descriptions and explanations of EvalML's features.

4.1 Automated Machine Learning (AutoML) Search

4.1.1 Background

Machine Learning

Machine learning (ML) is the process of constructing a mathematical model of a system based on a sample dataset collected from that system.

One of the main goals of training an ML model is to teach the model to separate the signal present in the data from the noise inherent in system and in the data collection process. If this is done effectively, the model can then be used to make accurate predictions about the system when presented with new, similar data. Additionally, introspecting on an ML model can reveal key information about the system being modeled, such as which inputs and transformations of the inputs are most useful to the ML model for learning the signal in the data, and are therefore the most predictive.

There are a **variety** of ML problem types. Supervised learning describes the case where the collected data contains an output value to be modeled and a set of inputs with which to train the model. EvalML focuses on training supervised learning models.

EvalML supports three common supervised ML problem types. The first is regression, where the target value to model is a continuous numeric value. Next are binary and multiclass classification, where the target value to model consists of two or more discrete values or categories. The choice of which supervised ML problem type is most appropriate depends on domain expertise and on how the model will be evaluated and used.

EvalML is currently building support for supervised time series problems: time series regression, time series binary classification, and time series multiclass classification. While we've added some features to tackle these kinds of problems, our functionality is still being actively developed so please be mindful of that before using it.

AutoML and Search

AutoML is the process of automating the construction, training and evaluation of ML models. Given a data and some configuration, AutoML searches for the most effective and accurate ML model or models to fit the dataset. During the search, AutoML will explore different combinations of model type, model parameters and model architecture.

An effective AutoML solution offers several advantages over constructing and tuning ML models by hand. AutoML can assist with many of the difficult aspects of ML, such as avoiding overfitting and underfitting, imbalanced data, detecting data leakage and other potential issues with the problem setup, and automatically applying best-practice data cleaning, feature engineering, feature selection and various modeling techniques. AutoML can also leverage

search algorithms to optimally sweep the hyperparameter search space, resulting in model performance which would be difficult to achieve by manual training.

4.1.2 AutoML in EvalML

EvalML supports all of the above and more.

In its simplest usage, the AutoML search interface requires only the input data, the target data and a `problem_type` specifying what kind of supervised ML problem to model.

** Graphing methods, like verbose `AutoMLSearch`, on Jupyter Notebook and Jupyter Lab require `ipywidgets` to be installed.

** If graphing on Jupyter Lab, `jupyterlab-plotly` required. To download this, make sure you have `npm` installed.

```
[1]: import evalml
from evalml.utils import infer_feature_types
X, y = evalml.demos.load_fraud(n_rows=250)
```

```

           Number of Features
Boolean                               1
Categorical                           6
Numeric                               5

Number of training examples: 250
Targets
False      88.40%
True       11.60%
Name: fraud, dtype: object
```

To provide data to EvalML, it is recommended that you initialize a `Woodwork` accessor on your data. This allows you to easily control how EvalML will treat each of your features before training a model.

EvalML also accepts `pandas` input, and will run type inference on top of the input `pandas` data. If you'd like to change the types inferred by EvalML, you can use the `infer_feature_types` utility method, which takes `pandas` or `numpy` input and converts it to a `Woodwork` data structure. The `feature_types` parameter can be used to specify what types specific columns should be.

Feature types such as `Natural Language` must be specified in this way, otherwise `Woodwork` will infer it as `Unknown` type and drop it during the `AutoMLSearch`.

In the example below, we reformat a couple features to make them easily consumable by the model, and then specify that the provider, which would have otherwise been inferred as a column with natural language, is a categorical column.

```
[2]: X.ww['expiration_date'] = X['expiration_date'].apply(lambda x: '20{}-01-{}'.format(x.
↳split("/") [1], x.split("/") [0]))
X = infer_feature_types(X, feature_types= {'store_id': 'categorical',
                                           'expiration_date': 'datetime',
                                           'lat': 'categorical',
                                           'lng': 'categorical',
                                           'provider': 'categorical'})
```

In order to validate the results of the pipeline creation and optimization process, we will save some of our data as a holdout set.

```
[3]: X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
↳problem_type='binary', test_size=.2)
```


Data Checks

Before calling `AutoMLSearch.search`, we should run some sanity checks on our data to ensure that the input data being passed will not run into some common issues before running a potentially time-consuming search. EvalML has various data checks that makes this easy. Each data check will return a collection of warnings and errors if it detects potential issues with the input data. This allows users to inspect their data to avoid confusing errors that may arise during the search process. You can learn about each of the data checks available through our [data checks guide](#)

Here, we will run the `DefaultDataChecks` class, which contains a series of data checks that are generally useful.

```
[4]: from evalml.data_checks import DefaultDataChecks

data_checks = DefaultDataChecks("binary", "log loss binary")
data_checks.validate(X_train, y_train)

[4]: {'warnings': [], 'errors': [], 'actions': []}
```

Since there were no warnings or errors returned, we can safely continue with the search process.

```
[5]: automl = evalml.automl.AutoMLSearch(X_train=X_train, y_train=y_train, problem_type=
↳ 'binary', verbose=True)
automl.search()

Using default limit of max_batches=1.

Generating pipelines to search over...
Removing columns ['currency'] because they are of 'Unknown' type
8 pipelines ready for search.

*****
* Beginning pipeline search *
*****

Optimizing for Log Loss Binary.
Lower score is better.

Using SequentialEngine to train and score pipelines.
Searching up to 1 batches for a total of 9 pipelines.
Allowed model families: linear_model, catboost, extra_trees, xgboost, decision_tree,
↳ random_forest, lightgbm

FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type': ...

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline
Mode Baseline Binary Classification Pipeline:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 3.970

*****
* Evaluating Batch Number 1 *
*****

Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization,
↳ Component + One Hot Encoder + Oversampler + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.627
```

(continues on next page)

(continued from previous page)

```

Decision Tree Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 5.404
    High coefficient of variation (cv >= 0.5) within cross validation scores.
    Decision Tree Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler may not perform as_
↳estimated on unseen data.
Random Forest Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.290
LightGBM Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.321
Logistic Regression Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.652
XGBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.274
Extra Trees Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.339
CatBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.627

Search finished after 00:17
Best pipeline: XGBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳Featurization Component + One Hot Encoder + Oversampler
Best pipeline Log Loss Binary: 0.273766

```

With the `verbose` argument set to `True`, the AutoML search will log its progress, reporting each pipeline and parameter set evaluated during the search.

There are a number of mechanisms to control the AutoML search time. One way is to set the `max_batches` parameter which controls the maximum number of rounds of AutoML to evaluate, where each round may train and score a variable number of pipelines. Another way is to set the `max_iterations` parameter which controls the maximum number of candidate models to be evaluated during AutoML. By default, AutoML will search for a single batch. The first pipeline to be evaluated will always be a baseline model representing a trivial solution.

The AutoML interface supports a variety of other parameters. For a comprehensive list, please [refer to the API reference](#).

We also provide [a standalone search method](#) which does all of the above in a single line, and returns the `AutoMLSearch` instance and data check results. If there were data check errors, AutoML will not be run and no `AutoMLSearch` instance will be returned.

Detecting Problem Type

EvalML includes a simple method, `detect_problem_type`, to help determine the problem type given the target data.

This function can return the predicted problem type as a `ProblemType` enum, choosing from `ProblemType.BINARY`, `ProblemType.MULTICLASS`, and `ProblemType.REGRESSION`. If the target data is invalid (for instance when there is only 1 unique label), the function will throw an error instead.

```
[6]: import pandas as pd
      from evalml.problem_types import detect_problem_type

      y_binary = pd.Series([0, 1, 1, 0, 1, 1])
      detect_problem_type(y_binary)

[6]: <ProblemTypes.BINARY: 'binary'>
```

Objective parameter

`AutoMLSearch` takes in an objective parameter to determine which objective to optimize for. By default, this parameter is set to `auto`, which allows `AutoML` to choose `LogLossBinary` for binary classification problems, `LogLossMulticlass` for multiclass classification problems, and `R2` for regression problems.

It should be noted that the objective parameter is only used in ranking and helping choose the pipelines to iterate over, but is not used to optimize each individual pipeline during fit-time.

To get the default objective for each problem type, you can use the `get_default_primary_search_objective` function.

```
[7]: from evalml automl import get_default_primary_search_objective

      binary_objective = get_default_primary_search_objective("binary")
      multiclass_objective = get_default_primary_search_objective("multiclass")
      regression_objective = get_default_primary_search_objective("regression")

      print(binary_objective.name)
      print(multiclass_objective.name)
      print(regression_objective.name)

      Log Loss Binary
      Log Loss Multiclass
      R2
```

Using custom pipelines

EvalML's `AutoML` algorithm generates a set of pipelines to search with. To provide a custom set instead, set `allowed_component_graphs` to a dictionary of custom component graphs. `AutoMLSearch` will use these to generate `Pipeline` instances. Note: this will prevent `AutoML` from generating other pipelines to search over.

```
[8]: from evalml.pipelines import MulticlassClassificationPipeline

      automl_custom = evalml.automl.AutoMLSearch(
          X_train=X_train,
          y_train=y_train,
          problem_type='multiclass',
```

(continues on next page)

(continued from previous page)

```
verbose=True,  
allowed_component_graphs={"My_pipeline": ['Simple Imputer', 'Random Forest_↵  
↵Classifier'],  
                           "My_other_pipeline": ['One Hot Encoder', 'Random Forest_↵  
↵Classifier']})
```

Using default limit of max_batches=1.

2 pipelines ready for search.

```
/home/docs/checkouts/readthedocs.org/user_builds/feature-labs-inc-evalml/envs/v0.32.1/  
↵lib/python3.8/site-packages/evalml/automl/automl_search.py:862:↵  
↵ParameterNotUsedWarning:
```

```
Parameters for components {'Oversampler'} will not be used to instantiate the_↵  
↵pipeline since they don't appear in the pipeline
```

Stopping the search early

To stop the search early, hit `Ctrl-C`. This will bring up a prompt asking for confirmation. Responding with `y` will immediately stop the search. Responding with `n` will continue the search.

Callback functions

`AutoMLSearch` supports several callback functions, which can be specified as parameters when initializing an `AutoMLSearch` object. They are:

- `start_iteration_callback`
- `add_result_callback`
- `error_callback`

Start Iteration Callback

Users can set `start_iteration_callback` to set what function is called before each pipeline training iteration. This callback function must take three positional parameters: the pipeline class, the pipeline parameters, and the `AutoMLSearch` object.

```
[9]: ## start_iteration_callback example function  
def start_iteration_callback_example(pipeline_class, pipeline_params, automl_obj):  
    print ("Training pipeline with the following parameters:", pipeline_params)
```

Add Result Callback

Users can set `add_result_callback` to set what function is called after each pipeline training iteration. This callback function must take three positional parameters: a dictionary containing the training results for the new pipeline, an `untrained_pipeline` containing the parameters used during training, and the `AutoMLSearch` object.

```
[10]: ## add_result_callback example function
def add_result_callback_example(pipeline_results_dict, untrained_pipeline, automl_
    obj):
    print ("Results for trained pipeline with the following parameters:", pipeline_
    results_dict)
```

Error Callback

Users can set the `error_callback` to set what function called when `search()` errors and raises an `Exception`. This callback function takes three positional parameters: the `Exception` raised, the `traceback`, and the `AutoMLSearch` object. This callback function must also accept `kwargs`, so `AutoMLSearch` is able to pass along other parameters used by default.

`Evalml` defines several error callback functions, which can be found under `evalml.automl.callbacks`. They are:

- `silent_error_callback`
- `raise_error_callback`
- `log_and_save_error_callback`
- `raise_and_save_error_callback`
- `log_error_callback` (default used when `error_callback` is `None`)

```
[11]: # error_callback example; this is implemented in the evalml library
def raise_error_callback(exception, traceback, automl, **kwargs):
    """Raises the exception thrown by the AutoMLSearch object. Also logs the_
    exception as an error."""
    logger.error(f'AutoMLSearch raised a fatal exception: {str(exception)}')
    logger.error("\n".join(traceback))
    raise exception
```

4.1.3 View Rankings

A summary of all the pipelines built can be returned as a `pandas DataFrame` which is sorted by score. The `score` column contains the average score across all cross-validation folds while the `validation_score` column is computed from the first cross-validation fold.

```
[12]: automl.rankings
```

```
[12]:   id  pipeline_name  search_order  \
0   6  XGBoost Classifier w/ Drop Columns Transformer...    6
1   3  Random Forest Classifier w/ Drop Columns Trans...    3
2   4  LightGBM Classifier w/ Drop Columns Transforme...    4
3   7  Extra Trees Classifier w/ Drop Columns Transfo...    7
4   8  CatBoost Classifier w/ Drop Columns Transforme...    8
5   1  Elastic Net Classifier w/ Drop Columns Transfo...    1
6   5  Logistic Regression Classifier w/ Drop Columns...    5
```

(continues on next page)

(continued from previous page)

7	0	Mode Baseline Binary Classification Pipeline			0
8	2	Decision Tree Classifier w/ Drop Columns Trans...			2
		mean_cv_score	standard_deviation_cv_score	validation_score	\
0		0.273766	0.164425	0.206398	
1		0.289832	0.045093	0.276903	
2		0.320864	0.147577	0.226765	
3		0.338676	0.030914	0.314494	
4		0.626968	0.005110	0.632822	
5		0.627427	0.108176	0.508699	
6		0.651602	0.117410	0.522408	
7		3.970423	0.266060	4.124033	
8		5.403917	2.627562	2.632694	
		percent_better_than_baseline	high_variance_cv	\	
0		93.104871	False		
1		92.700224	False		
2		91.918642	False		
3		91.470030	False		
4		84.209030	False		
5		84.197481	False		
6		83.588598	False		
7		0.000000	False		
8		-36.104299	True		
		parameters			
0		{ 'Drop Columns Transformer': { 'columns': ['cur...			
1		{ 'Drop Columns Transformer': { 'columns': ['cur...			
2		{ 'Drop Columns Transformer': { 'columns': ['cur...			
3		{ 'Drop Columns Transformer': { 'columns': ['cur...			
4		{ 'Drop Columns Transformer': { 'columns': ['cur...			
5		{ 'Drop Columns Transformer': { 'columns': ['cur...			
6		{ 'Drop Columns Transformer': { 'columns': ['cur...			
7		{ 'Baseline Classifier': { 'strategy': 'mode' }}			
8		{ 'Drop Columns Transformer': { 'columns': ['cur...			

4.1.4 Describe Pipeline

Each pipeline is given an `id`. We can get more information about any particular pipeline using that `id`. Here, we will get more information about the pipeline with `id = 1`.

```
[13]: automl.describe_pipeline(1)
```

```
*****
* Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳ Featurization Component + One Hot Encoder + Oversampler + Standard Scaler *
*****

Problem Type: binary
Model Family: Linear

Pipeline Steps
=====
1. Drop Columns Transformer
```

(continues on next page)

(continued from previous page)

```

    * columns : ['currency']
2. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
3. DateTime Featurization Component
    * features_to_extract : ['year', 'month', 'day_of_week', 'hour']
    * encode_as_categories : False
    * date_index : None
4. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
5. Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * k_neighbors : 5
6. Standard Scaler
7. Elastic Net Classifier
    * penalty : elasticnet
    * C : 1.0
    * l1_ratio : 0.15
    * n_jobs : -1
    * multi_class : auto
    * solver : saga

Training
=====
Training for binary problems.
Total training time (including CV): 1.9 seconds

Cross Validation
-----

```

	Log Loss Binary	MCC Binary	Gini	AUC	Precision	F1	Balanced_
↪Accuracy Binary	Accuracy Binary	# Training	# Validation				
0	0.509	0.046	0.174	0.587	0.167	0.143	↪
↪ 0.520	0.821	133		67			
1	0.653	0.290	0.131	0.566	0.375	0.375	↪
↪ 0.645	0.851	133		67			
2	0.720	0.062	0.080	0.540	0.143	0.190	↪
↪ 0.541	0.742	134		66			
mean	0.627	0.133	0.128	0.564	0.228	0.236	↪
↪ 0.569	0.805	-		-			
std	0.108	0.137	0.047	0.023	0.128	0.123	↪
↪ 0.067	0.056	-		-			
coef of var	0.172	1.031	0.366	0.042	0.560	0.519	↪
↪ 0.118	0.070	-		-			

4.1.5 Get Pipeline

We can get the object of any pipeline via their id as well:

```
[14]: pipeline = automl.get_pipeline(1)
print(pipeline.name)
print(pipeline.parameters)

Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler + Standard Scaler
{'Drop Columns Transformer': {'columns': ['currency']}, 'Imputer': {'categorical_
↳impute_strategy': 'most_frequent', 'numeric_impute_strategy': 'mean', 'categorical_
↳fill_value': None, 'numeric_fill_value': None}, 'DateTime Featurization Component':
↳{'features_to_extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_
↳categories': False, 'date_index': None}, 'One Hot Encoder': {'top_n': 10, 'features_
↳to_encode': None, 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore
↳', 'handle_missing': 'error'}, 'Oversampler': {'sampling_ratio': 0.25, 'k_neighbors_
↳default': 5, 'n_jobs': -1, 'sampling_ratio_dict': None, 'k_neighbors': 5}, 'Elastic_
↳Net Classifier': {'penalty': 'elasticnet', 'C': 1.0, 'l1_ratio': 0.15, 'n_jobs': -1,
↳ 'multi_class': 'auto', 'solver': 'saga'}}
```

Get best pipeline

If you specifically want to get the best pipeline, there is a convenient accessor for that. The pipeline returned is already fitted on the input X, y data that we passed to AutoMLSearch. To turn off this default behavior, set `train_best_pipeline=False` when initializing AutoMLSearch.

```
[15]: best_pipeline = automl.best_pipeline
print(best_pipeline.name)
print(best_pipeline.parameters)
best_pipeline.predict(X_train)

XGBoost Classifier w/ Drop Columns Transformer + Imputer + DateTime Featurization_
↳Component + One Hot Encoder + Oversampler
{'Drop Columns Transformer': {'columns': ['currency']}, 'Imputer': {'categorical_
↳impute_strategy': 'most_frequent', 'numeric_impute_strategy': 'mean', 'categorical_
↳fill_value': None, 'numeric_fill_value': None}, 'DateTime Featurization Component':
↳{'features_to_extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_
↳categories': False, 'date_index': None}, 'One Hot Encoder': {'top_n': 10, 'features_
↳to_encode': None, 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore
↳', 'handle_missing': 'error'}, 'Oversampler': {'sampling_ratio': 0.25, 'k_neighbors_
↳default': 5, 'n_jobs': -1, 'sampling_ratio_dict': None, 'k_neighbors': 5}, 'XGBoost_
↳Classifier': {'eta': 0.1, 'max_depth': 6, 'min_child_weight': 1, 'n_estimators':
↳100, 'n_jobs': -1, 'eval_metric': 'logloss'}}
```

```
[15]: 0      False
      1      False
      2      False
      3      False
      4      False
      ...
      195    False
      196    False
      197    False
      198    False
      199    False
      Name: fraud, Length: 200, dtype: bool
```


4.1.6 Training and Scoring Multiple Pipelines using AutoMLSearch

AutoMLSearch will automatically fit the best pipeline on the entire training data. It also provides an easy API for training and scoring other pipelines.

If you'd like to train one or more pipelines on the entire training data, you can use the `train_pipelines` method

Similarly, if you'd like to score one or more pipelines on a particular dataset, you can use the `train_pipelines` method

```
[16]: trained_pipelines = automl.train_pipelines([automl.get_pipeline(i) for i in [0, 1, 2]])
trained_pipelines

[16]: {'Mode Baseline Binary Classification Pipeline': pipeline =
↳ BinaryClassificationPipeline(component_graph={'Baseline Classifier': ['Baseline_
↳ Classifier', 'X', 'y']}, parameters={'Baseline Classifier':{'strategy': 'mode'}},
↳ custom_name='Mode Baseline Binary Classification Pipeline', random_seed=0),
' Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳ Featurization Component + One Hot Encoder + Oversampler + Standard Scaler':
↳ pipeline = BinaryClassificationPipeline(component_graph={'Drop Columns Transformer':
↳ ['Drop Columns Transformer', 'X', 'y'], 'Imputer': ['Imputer', 'Drop Columns_
↳ Transformer.x', 'y'], 'DateTime Featurization Component': ['DateTime Featurization_
↳ Component', 'Imputer.x', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'DateTime_
↳ Featurization Component.x', 'y'], 'Oversampler': ['Oversampler', 'One Hot Encoder.x
↳ ', 'y'], 'Standard Scaler': ['Standard Scaler', 'Oversampler.x', 'Oversampler.y'],
↳ 'Elastic Net Classifier': ['Elastic Net Classifier', 'Standard Scaler.x',
↳ 'Oversampler.y']}, parameters={'Drop Columns Transformer':{'columns': ['currency']},
↳ 'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_impute_strategy
↳ ': 'mean', 'categorical_fill_value': None, 'numeric_fill_value': None}, 'DateTime_
↳ Featurization Component':{'features_to_extract': ['year', 'month', 'day_of_week',
↳ 'hour'], 'encode_as_categories': False, 'date_index': None}, 'One Hot Encoder':{'
↳ 'top_n': 10, 'features_to_encode': None, 'categories': None, 'drop': 'if_binary',
↳ 'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'Oversampler':{'sampling_
↳ ratio': 0.25, 'k_neighbors_default': 5, 'n_jobs': -1, 'sampling_ratio_dict': None,
↳ 'k_neighbors': 5}, 'Elastic Net Classifier':{'penalty': 'elasticnet', 'C': 1.0, 'l1_
↳ ratio': 0.15, 'n_jobs': -1, 'multi_class': 'auto', 'solver': 'saga'}}}, random_
↳ seed=0),
' Decision Tree Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳ Featurization Component + One Hot Encoder + Oversampler': pipeline =
↳ BinaryClassificationPipeline(component_graph={'Drop Columns Transformer': ['Drop_
↳ Columns Transformer', 'X', 'y'], 'Imputer': ['Imputer', 'Drop Columns Transformer.x
↳ ', 'y'], 'DateTime Featurization Component': ['DateTime Featurization Component',
↳ 'Imputer.x', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'DateTime Featurization_
↳ Component.x', 'y'], 'Oversampler': ['Oversampler', 'One Hot Encoder.x', 'y'],
↳ 'Decision Tree Classifier': ['Decision Tree Classifier', 'Oversampler.x',
↳ 'Oversampler.y']}, parameters={'Drop Columns Transformer':{'columns': ['currency']},
↳ 'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_impute_strategy
↳ ': 'mean', 'categorical_fill_value': None, 'numeric_fill_value': None}, 'DateTime_
↳ Featurization Component':{'features_to_extract': ['year', 'month', 'day_of_week',
↳ 'hour'], 'encode_as_categories': False, 'date_index': None}, 'One Hot Encoder':{'
↳ 'top_n': 10, 'features_to_encode': None, 'categories': None, 'drop': 'if_binary',
↳ 'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'Oversampler':{'sampling_
↳ ratio': 0.25, 'k_neighbors_default': 5, 'n_jobs': -1, 'sampling_ratio_dict': None,
↳ 'k_neighbors': 5}, 'Decision Tree Classifier':{'criterion': 'gini', 'max_features':
↳ 'auto', 'max_depth': 6, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0}},
↳ random_seed=0)}
```

```
[17]: pipeline_holdout_scores = automl.score_pipelines([trained_pipelines[name] for name in
↳ trained_pipelines.keys()],
                                                    X_holdout,
                                                    y_holdout,
                                                    ['Accuracy Binary', 'F1', 'AUC'])
pipeline_holdout_scores

[17]: {'Mode Baseline Binary Classification Pipeline': OrderedDict([('Accuracy Binary',
0.88),
('F1', 0.0),
('AUC', 0.5)]),
'Elastic Net Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳ Featurization Component + One Hot Encoder + Oversampler + Standard Scaler':
↳ OrderedDict([('Accuracy Binary',
0.44),
('F1', 0.17647058823529413),
('AUC', 0.49242424242424243)]),
'Decision Tree Classifier w/ Drop Columns Transformer + Imputer + DateTime_
↳ Featurization Component + One Hot Encoder + Oversampler': OrderedDict([('Accuracy_
↳ Binary',
0.42),
('F1', 0.17142857142857143),
('AUC', 0.48295454545454547)])}
```

4.1.7 Saving AutoMLSearch and pipelines from AutoMLSearch

There are two ways to save results from AutoMLSearch.

- You can save the AutoMLSearch object itself, calling `.save(<filepath>)` to do so. This will allow you to save the AutoMLSearch state and reload all pipelines from this.
- If you want to save a pipeline from AutoMLSearch for future use, pipeline classes themselves have a `.save(<filepath>)` method.

```
[18]: # saving the entire automl search
automl.save("automl.cloudpickle")
automl2 = evalml.automl.AutoMLSearch.load("automl.cloudpickle")
# saving the best pipeline using .save()
best_pipeline.save("pipeline.cloudpickle")
best_pipeline_copy = evalml.pipelines.PipelineBase.load("pipeline.cloudpickle")
```

4.1.8 Limiting the AutoML Search Space

The AutoML search algorithm first trains each component in the pipeline with their default values. After the first iteration, it then tweaks the parameters of these components using the pre-defined hyperparameter ranges that these components have. To limit the search over certain hyperparameter ranges, you can specify a `custom_hyperparameters` argument with your AutoMLSearch parameters. These parameters will limit the hyperparameter search space.

Hyperparameter ranges can be found through the [API reference](#) for each component. Parameter arguments must be specified as dictionaries, but the associated values can be single values or `skopt.space` Real, Integer, Categorical values.

If however you'd like to specify certain values for the initial batch of the AutoML search algorithm, you can use the `pipeline_parameters` argument. This will set the initial batch's component parameters to the values passed by this argument.

```
[19]: from evalml import AutoMLSearch
from evalml.demos import load_fraud
from skopt.space import Categorical
from evalml.model_family import ModelFamily
import woodwork as ww

X, y = load_fraud(n_rows=1000)

# example of setting parameter to just one value
custom_hyperparameters = {'Imputer': {
    'numeric_impute_strategy': 'mean'
}}

# limit the numeric impute strategy to include only `median` and `most_frequent`
# `mean` is the default value for this argument, but it doesn't need to be included,
# → in the specified hyperparameter range for this to work
custom_hyperparameters = {'Imputer': {
    'numeric_impute_strategy': Categorical(['median', 'most_frequent'])
}}
# set the initial batch numeric impute strategy to 'median'
pipeline_parameters = {'Imputer': {
    'numeric_impute_strategy': 'median'
}}

# using this custom hyperparameter means that our Imputer components in these
# → pipelines will only search through
# 'median' and 'most_frequent' strategies for 'numeric_impute_strategy', and the
# → initial batch parameter will be
# set to 'median'
automl_constrained = AutoMLSearch(X_train=X, y_train=y, problem_type='binary',
                                  pipeline_parameters=pipeline_parameters,
                                  custom_hyperparameters=custom_hyperparameters,
                                  verbose=True)

      Number of Features
Boolean                1
Categorical            6
Numeric                5

Number of training examples: 1000
Targets
False      85.90%
True       14.10%
Name: fraud, dtype: object
Using default limit of max_batches=1.

Generating pipelines to search over...
8 pipelines ready for search.
```

4.1.9 Imbalanced Data

The AutoML search algorithm now has functionality to handle imbalanced data during classification! AutoMLSearch now provides two additional parameters, `sampler_method` and `sampler_balanced_ratio`, that allow you to let AutoMLSearch know whether to sample imbalanced data, and how to do so. `sampler_method` takes in either `Undersampler`, `Oversampler`, `auto`, or `None` as the sampler to use, and `sampler_balanced_ratio` specifies the minority/majority ratio that you want to sample to. Details on the `Undersampler` and `Oversampler` components can be found in the [documentation](#).

This can be used for imbalanced datasets, like the fraud dataset, which has a ‘minority:majority’ ratio of < 0.2 .

```
[20]: automl_auto = AutoMLSearch(X_train=X, y_train=y, problem_type='binary')
      automl_auto.allowed_pipelines[-1]

[20]: pipeline = BinaryClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X',
↪ 'y'], 'DateTime Featurization Component': ['DateTime Featurization Component',
↪ 'Imputer.x', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'DateTime Featurization_
↪ Component.x', 'y'], 'Oversampler': ['Oversampler', 'One Hot Encoder.x', 'y'],
↪ 'Standard Scaler': ['Standard Scaler', 'Oversampler.x', 'Oversampler.y'], 'Logistic_
↪ Regression Classifier': ['Logistic Regression Classifier', 'Standard Scaler.x',
↪ 'Oversampler.y']}, parameters={'Imputer':{'categorical_impute_strategy': 'most_
↪ frequent', 'numeric_impute_strategy': 'mean', 'categorical_fill_value': None,
↪ 'numeric_fill_value': None}, 'DateTime Featurization Component':{'features_to_
↪ extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_categories': False,
↪ 'date_index': None}, 'One Hot Encoder':{'top_n': 10, 'features_to_encode': None,
↪ 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore', 'handle_missing
↪ ': 'error'}, 'Oversampler':{'sampling_ratio': 0.25, 'k_neighbors_default': 5, 'n_
↪ jobs': -1, 'sampling_ratio_dict': None}, 'Logistic Regression Classifier':{'penalty
↪ ': 'l2', 'C': 1.0, 'n_jobs': -1, 'multi_class': 'auto', 'solver': 'lbfgs'}}, random_
↪ seed=0)
```

The `Oversampler` is chosen as the default sampling component here, since the `sampler_balanced_ratio = 0.25`. If you specified a lower ratio, for instance `sampler_balanced_ratio = 0.1`, then there would be no sampling component added here. This is because if a ratio of 0.1 would be considered balanced, then a ratio of 0.2 would also be balanced.

The `Oversampler` uses SMOTE under the hood, and automatically selects whether to use SMOTE, SMOTEN, or SMOTENC based on the data it receives.

```
[21]: automl_auto_ratio = AutoMLSearch(X_train=X, y_train=y, problem_type='binary', sampler_
↪ balanced_ratio=0.1)
      automl_auto_ratio.allowed_pipelines[-1]

[21]: pipeline = BinaryClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X',
↪ 'y'], 'DateTime Featurization Component': ['DateTime Featurization Component',
↪ 'Imputer.x', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'DateTime Featurization_
↪ Component.x', 'y'], 'Standard Scaler': ['Standard Scaler', 'One Hot Encoder.x', 'y
↪ '], 'Logistic Regression Classifier': ['Logistic Regression Classifier', 'Standard_
↪ Scaler.x', 'y']}, parameters={'Imputer':{'categorical_impute_strategy': 'most_
↪ frequent', 'numeric_impute_strategy': 'mean', 'categorical_fill_value': None,
↪ 'numeric_fill_value': None}, 'DateTime Featurization Component':{'features_to_
↪ extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_categories': False,
↪ 'date_index': None}, 'One Hot Encoder':{'top_n': 10, 'features_to_encode': None,
↪ 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore', 'handle_missing
↪ ': 'error'}, 'Logistic Regression Classifier':{'penalty': 'l2', 'C': 1.0, 'n_jobs': -
↪ 1, 'multi_class': 'auto', 'solver': 'lbfgs'}}, random_seed=0)
```

Additionally, you can add more fine-grained sampling ratios by passing in a `sampling_ratio_dict` in pipeline parameters. For this dictionary, AutoMLSearch expects the keys to be int values from 0 to $n-1$ for the classes, and the

values would be the `sampler_balanced_ratio` associated with each target. This dictionary would override the AutoML argument `sampler_balanced_ratio`. Below, you can see the scenario for Oversampler component on this dataset. Note that the logic for Undersamplers is included in the commented section.

```
[22]: # In this case, the majority class is the negative class
# for the oversampler, we don't want to oversample this class, so class 0 (majority)
# will have a ratio of 1 to itself
# for the minority class 1, we want to oversample it to have a minority/majority
# ratio of 0.5, which means we want minority to have 1/2 the samples as the minority
sampler_ratio_dict = {0: 1, 1: 0.5}
pipeline_parameters = {"Oversampler": {"sampler_balanced_ratio": sampler_ratio_dict}}
automl_auto_ratio_dict = AutoMLSearch(X_train=X, y_train=y, problem_type='binary',
# pipeline_parameters=pipeline_parameters)
automl_auto_ratio_dict.allowed_pipelines[-1]

# Undersampler case
# we don't want to undersample this class, so class 1 (minority) will have a ratio of
# 1 to itself
# for the majority class 0, we want to undersample it to have a minority/majority
# ratio of 0.5, which means we want majority to have 2x the samples as the minority
# sampler_ratio_dict = {0: 0.5, 1: 1}
# pipeline_parameters = {"Oversampler": {"sampler_balanced_ratio": sampler_ratio_dict}
# }
# automl_auto_ratio_dict = AutoMLSearch(X_train=X, y_train=y, problem_type='binary',
# pipeline_parameters=pipeline_parameters)

[22]: pipeline = BinaryClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X',
# 'y'], 'DateTime Featurization Component': ['DateTime Featurization Component',
# 'Imputer.x', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'DateTime Featurization
# Component.x', 'y'], 'Oversampler': ['Oversampler', 'One Hot Encoder.x', 'y'],
# 'Standard Scaler': ['Standard Scaler', 'Oversampler.x', 'Oversampler.y'], 'Logistic
# Regression Classifier': ['Logistic Regression Classifier', 'Standard Scaler.x',
# 'Oversampler.y']}, parameters={'Imputer':{'categorical_impute_strategy': 'most_
# frequent', 'numeric_impute_strategy': 'mean', 'categorical_fill_value': None,
# 'numeric_fill_value': None}, 'DateTime Featurization Component':{'features_to_
# extract': ['year', 'month', 'day_of_week', 'hour'], 'encode_as_categories': False,
# 'date_index': None}, 'One Hot Encoder':{'top_n': 10, 'features_to_encode': None,
# 'categories': None, 'drop': 'if_binary', 'handle_unknown': 'ignore', 'handle_missing
# ': 'error'}, 'Oversampler':{'sampling_ratio': 0.25, 'k_neighbors_default': 5, 'n_
# jobs': -1, 'sampling_ratio_dict': None, 'sampler_balanced_ratio': {0: 1, 1: 0.5}},
# 'Logistic Regression Classifier':{'penalty': 'l2', 'C': 1.0, 'n_jobs': -1, 'multi_
# class': 'auto', 'solver': 'lbfgs'}}}, random_seed=0)
```

4.1.10 Adding ensemble methods to AutoML

Stacking

Stacking is an ensemble machine learning algorithm that involves training a model to best combine the predictions of several base learning algorithms. First, each base learning algorithms is trained using the given data. Then, the combining algorithm or meta-learner is trained on the predictions made by those base learning algorithms to make a final prediction.

AutoML enables stacking using the `ensembling` flag during initialization; this is set to `False` by default. The stacking ensemble pipeline runs in its own batch after a whole cycle of training has occurred (each allowed pipeline trains for one batch). Note that this means **a large number of iterations may need to run before the stacking**

ensemble runs. It is also important to note that **only the first CV fold is calculated for stacking ensembles** because the model internally uses CV folds.

```
[23]: X, y = evalml.demos.load_breast_cancer()

automl_with_ensembling = AutoMLSearch(X_train=X, y_train=y,
                                     problem_type="binary",
                                     allowed_model_families=[ModelFamily.LINEAR_
↪MODEL],
                                     max_batches=4,
                                     ensembling=True,
                                     verbose=True)

automl_with_ensembling.search()
```

```
      Number of Features
Numeric                30
```

Number of training examples: 569

Targets

```
benign          62.74%
```

```
malignant      37.26%
```

Name: target, dtype: object

Generating pipelines to search over...

2 pipelines ready for search.

Ensembling will run every 3 batches.

```
*****
* Beginning pipeline search *
*****
```

Optimizing for Log Loss Binary.

Lower score is better.

Using SequentialEngine to train and score pipelines.

Searching up to 4 batches for a total of 14 pipelines.

Allowed model families: linear_model

```
FigureWidget({
  'data': [{'mode': 'lines+markers',
            'name': 'Best Score',
            'type'...
```

Evaluating Baseline Pipeline: Mode Baseline Binary Classification Pipeline

Mode Baseline Binary Classification Pipeline:

Starting cross validation

Finished cross validation - mean Log Loss Binary: 12.868

```
*****
* Evaluating Batch Number 1 *
*****
```

Elastic Net Classifier w/ Imputer + Standard Scaler:

Starting cross validation

Finished cross validation - mean Log Loss Binary: 0.077

Logistic Regression Classifier w/ Imputer + Standard Scaler:

Starting cross validation

Finished cross validation - mean Log Loss Binary: 0.077

(continues on next page)

(continued from previous page)

```

*****
* Evaluating Batch Number 2 *
*****

Logistic Regression Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.097
Logistic Regression Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.085
Logistic Regression Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.097
Logistic Regression Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.091
Logistic Regression Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.080

*****
* Evaluating Batch Number 3 *
*****

Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.075
Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.075
Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.079
Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.076
Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.075

*****
* Evaluating Batch Number 4 *
*****

Sklearn Stacked Ensemble Classification Pipeline:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.123

Search finished after 00:24
Best pipeline: Elastic Net Classifier w/ Imputer + Standard Scaler
Best pipeline Log Loss Binary: 0.075387

```

We can view more information about the stacking ensemble pipeline (which was the best performing pipeline) by calling `.describe()`.

```
[24]: automl_with_ensembling.best_pipeline.describe()
```

```

*****
* Elastic Net Classifier w/ Imputer + Standard Scaler *
*****

Problem Type: binary
Model Family: Linear
Number of features: 30

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : median
    * categorical_fill_value : None
    * numeric_fill_value : None
2. Standard Scaler
3. Elastic Net Classifier
    * penalty : elasticnet
    * C : 8.123565600467177
    * l1_ratio : 0.47997717237505744
    * n_jobs : -1
    * multi_class : auto
    * solver : saga

```

4.1.11 Access raw results

The `AutoMLSearch` class records detailed results information under the `results` field, including information about the cross-validation scoring and parameters.

```
[25]: automl.results
```

```

[25]: {'pipeline_results': {0: {'id': 0,
    'pipeline_name': 'Mode Baseline Binary Classification Pipeline',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
    ↪BinaryClassificationPipeline,
    'pipeline_summary': 'Baseline Classifier',
    'parameters': {'Baseline Classifier': {'strategy': 'mode'}},
    'mean_cv_score': 3.970423187263591,
    'standard_deviation_cv_score': 0.26606000431837074,
    'high_variance_cv': False,
    'training_time': 0.914386510848999,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
        4.124033002377396),
        ('MCC Binary', 0.0),
        ('Gini', 0.0),
        ('AUC', 0.5),
        ('Precision', 0.0),
        ('F1', 0.0),
        ('Balanced Accuracy Binary', 0.5),
        ('Accuracy Binary', 0.8805970149253731),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 4.124033002377396,
    'binary_classification_threshold': 9.16384630183206e-53},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',

```

(continues on next page)

(continued from previous page)

```

        4.124033002377395),
        ('MCC Binary', 0.0),
        ('Gini', 0.0),
        ('AUC', 0.5),
        ('Precision', 0.0),
        ('F1', 0.0),
        ('Balanced Accuracy Binary', 0.5),
        ('Accuracy Binary', 0.8805970149253731),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 4.124033002377395,
    'binary_classification_threshold': 9.16384630183206e-53},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
        3.6632035570359824),
        ('MCC Binary', 0.0),
        ('Gini', 0.0),
        ('AUC', 0.5),
        ('Precision', 0.0),
        ('F1', 0.0),
        ('Balanced Accuracy Binary', 0.5),
        ('Accuracy Binary', 0.8939393939393939),
        ('# Training', 134),
        ('# Validation', 66)]),
    'mean_cv_score': 3.6632035570359824,
    'binary_classification_threshold': 9.16384630183206e-53}},
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 0,
'MCC Binary': 0,
'Gini': 0,
'AUC': 0,
'Precision': 0,
'F1': 0,
'Balanced Accuracy Binary': 0,
'Accuracy Binary': 0},
'percent_better_than_baseline': 0,
'validation_score': 4.124033002377396},
1: {'id': 1,
    'pipeline_name': 'Elastic Net Classifier w/ Drop Columns Transformer + Imputer +
↳ DateTime Featurization Component + One Hot Encoder + Oversampler + Standard Scaler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↳ BinaryClassificationPipeline,
    'pipeline_summary': 'Elastic Net Classifier w/ Drop Columns Transformer + Imputer_
↳ + DateTime Featurization Component + One Hot Encoder + Oversampler + Standard Scaler
↳ ',
    'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
    'Imputer': {'categorical_impute_strategy': 'most_frequent',
    'numeric_impute_strategy': 'mean',
    'categorical_fill_value': None,
    'numeric_fill_value': None},
    'DateTime Featurization Component': {'features_to_extract': ['year',
    'month',
    'day_of_week',
    'hour']},
    'encode_as_categories': False,
    'date_index': None},
    'One Hot Encoder': {'top_n': 10,
    'features_to_encode': None,
    'categories': None,

```

(continues on next page)

(continued from previous page)

```

'drop': 'if_binary',
'handle_unknown': 'ignore',
'handle_missing': 'error'},
'Oversampler': {'sampling_ratio': 0.25,
'k_neighbors_default': 5,
'n_jobs': -1,
'sampling_ratio_dict': None,
'k_neighbors': 5},
'Elastic Net Classifier': {'penalty': 'elasticnet',
'C': 1.0,
'1l_ratio': 0.15,
'n_jobs': -1,
'multi_class': 'auto',
'solver': 'saga'}},
'mean_cv_score': 0.6274268949098596,
'standard_deviation_cv_score': 0.10817615334446118,
'high_variance_cv': False,
'training_time': 1.9461102485656738,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.5086986589950817),
('MCC Binary', 0.045713245813732276),
('Gini', 0.17372881355932202),
('AUC', 0.586864406779661),
('Precision', 0.16666666666666666),
('F1', 0.14285714285714288),
('Balanced Accuracy Binary', 0.5201271186440678),
('Accuracy Binary', 0.8208955223880597),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.5086986589950817,
'binary_classification_threshold': 0.3735112604474508},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.6531787435325545),
('MCC Binary', 0.2902542372881356),
('Gini', 0.13135593220338992),
('AUC', 0.565677966101695),
('Precision', 0.375),
('F1', 0.375),
('Balanced Accuracy Binary', 0.6451271186440678),
('Accuracy Binary', 0.8507462686567164),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.6531787435325545,
'binary_classification_threshold': 0.23238767434564783},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.7204032822019426),
('MCC Binary', 0.0620066404299915),
('Gini', 0.07990314769975781),
('AUC', 0.5399515738498789),
('Precision', 0.14285714285714285),
('F1', 0.19047619047619047),
('Balanced Accuracy Binary', 0.5411622276029056),
('Accuracy Binary', 0.7424242424242424),
('# Training', 134),
('# Validation', 66)]),
'mean_cv_score': 0.7204032822019426,
'binary_classification_threshold': 0.45539243933617346}],

```

(continues on next page)

(continued from previous page)

```

    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 84.
↪19748058789973,
    'MCC Binary': inf,
    'Gini': inf,
    'AUC': 6.416464891041162,
    'Precision': 22.817460317460316,
    'F1': 23.611111111111114,
    'Balanced Accuracy Binary': 6.880548829701372,
    'Accuracy Binary': -8.035579677370707},
    'percent_better_than_baseline': 84.19748058789973,
    'validation_score': 0.5086986589950817},
2: {'id': 2,
    'pipeline_name': 'Decision Tree Classifier w/ Drop Columns Transformer + Imputer +
↪DateTime Featurization Component + One Hot Encoder + Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'Decision Tree Classifier w/ Drop Columns Transformer +
↪Imputer + DateTime Featurization Component + One Hot Encoder + Oversampler',
    'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
    'Imputer': {'categorical_impute_strategy': 'most_frequent',
    'numeric_impute_strategy': 'mean',
    'categorical_fill_value': None,
    'numeric_fill_value': None},
    'DateTime Featurization Component': {'features_to_extract': ['year',
    'month',
    'day_of_week',
    'hour'],
    'encode_as_categories': False,
    'date_index': None},
    'One Hot Encoder': {'top_n': 10,
    'features_to_encode': None,
    'categories': None,
    'drop': 'if_binary',
    'handle_unknown': 'ignore',
    'handle_missing': 'error'},
    'Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'k_neighbors': 5},
    'Decision Tree Classifier': {'criterion': 'gini',
    'max_features': 'auto',
    'max_depth': 6,
    'min_samples_split': 2,
    'min_weight_fraction_leaf': 0.0}},
    'mean_cv_score': 5.4039166268422525,
    'standard_deviation_cv_score': 2.627561787321521,
    'high_variance_cv': True,
    'training_time': 1.832096815109253,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
    2.632694044787486),
    ('MCC Binary', 0.0),
    ('Gini', 0.22669491525423724),
    ('AUC', 0.6133474576271186),
    ('Precision', 0.11940298507462686),
    ('F1', 0.21333333333333335),
    ('Balanced Accuracy Binary', 0.5),

```

(continues on next page)

(continued from previous page)

```

        ('Accuracy Binary', 0.11940298507462686),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 2.632694044787486,
    'binary_classification_threshold': -1.4827414887138564e-52},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
        5.719796644733923),
        ('MCC Binary', 0.17518582316850065),
        ('Gini', 0.12711864406779672),
        ('AUC', 0.5635593220338984),
        ('Precision', 0.2857142857142857),
        ('F1', 0.26666666666666666),
        ('Balanced Accuracy Binary', 0.5826271186440678),
        ('Accuracy Binary', 0.835820895522388),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 5.719796644733923,
    'binary_classification_threshold': 0.9999999885588493},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
        7.859259191005347),
        ('MCC Binary', -0.008317464424851172),
        ('Gini', -0.09685230024213076),
        ('AUC', 0.4515738498789346),
        ('Precision', 0.1),
        ('F1', 0.11764705882352941),
        ('Balanced Accuracy Binary', 0.4951573849878934),
        ('Accuracy Binary', 0.7727272727272727),
        ('# Training', 134),
        ('# Validation', 66)]),
    'mean_cv_score': 7.859259191005347,
    'binary_classification_threshold': 0.9999999885588493}],
    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': -36.
    ↪104298508457546,
    'MCC Binary': inf,
    'Gini': inf,
    'AUC': 4.2826876513317185,
    'Precision': 16.83724235963042,
    'F1': 19.92156862745098,
    'Balanced Accuracy Binary': 2.592816787732044,
    'Accuracy Binary': -30.906075682195077},
    'percent_better_than_baseline': -36.104298508457546,
    'validation_score': 2.632694044787486},
    3: {'id': 3,
    'pipeline_name': 'Random Forest Classifier w/ Drop Columns Transformer + Imputer +
    ↪DateTime Featurization Component + One Hot Encoder + Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
    ↪BinaryClassificationPipeline,
    'pipeline_summary': 'Random Forest Classifier w/ Drop Columns Transformer +
    ↪Imputer + DateTime Featurization Component + One Hot Encoder + Oversampler',
    'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
    'Imputer': {'categorical_impute_strategy': 'most_frequent',
    'numeric_impute_strategy': 'mean',
    'categorical_fill_value': None,
    'numeric_fill_value': None},
    'DateTime Featurization Component': {'features_to_extract': ['year',
    'month',
    'day_of_week',

```

(continues on next page)

(continued from previous page)

```

    'hour'],
    'encode_as_categories': False,
    'date_index': None},
    'One Hot Encoder': {'top_n': 10,
    'features_to_encode': None,
    'categories': None,
    'drop': 'if_binary',
    'handle_unknown': 'ignore',
    'handle_missing': 'error'},
    'Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'k_neighbors': 5},
    'Random Forest Classifier': {'n_estimators': 100,
    'max_depth': 6,
    'n_jobs': -1}},
    'mean_cv_score': 0.2898320159158559,
    'standard_deviation_cv_score': 0.04509254687158935,
    'high_variance_cv': False,
    'training_time': 2.0981149673461914,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.27690265474464915),
    ('MCC Binary', 0.47636443708895493),
    ('Gini', 0.5932203389830508),
    ('AUC', 0.7966101694915254),
    ('Precision', 1.0),
    ('F1', 0.4),
    ('Balanced Accuracy Binary', 0.625),
    ('Accuracy Binary', 0.9104477611940298),
    ('# Training', 133),
    ('# Validation', 67)]),
    'mean_cv_score': 0.27690265474464915,
    'binary_classification_threshold': 0.5255549360104743},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.33997691511520767),
    ('MCC Binary', 0.36540212737375455),
    ('Gini', 0.31779661016949157),
    ('AUC', 0.6588983050847458),
    ('Precision', 0.6666666666666666),
    ('F1', 0.36363636363636365),
    ('Balanced Accuracy Binary', 0.6165254237288136),
    ('Accuracy Binary', 0.8955223880597015),
    ('# Training', 133),
    ('# Validation', 67)]),
    'mean_cv_score': 0.33997691511520767,
    'binary_classification_threshold': 0.4861948486476305},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.2526164778877109),
    ('MCC Binary', 0.5132181379714255),
    ('Gini', 0.7239709443099271),
    ('AUC', 0.8619854721549636),
    ('Precision', 1.0),
    ('F1', 0.44444444444444445),
    ('Balanced Accuracy Binary', 0.6428571428571428),
    ('Accuracy Binary', 0.9242424242424242),
    ('# Training', 134),

```

(continues on next page)

(continued from previous page)

```

        ('# Validation', 66)]),
        'mean_cv_score': 0.2526164778877109,
        'binary_classification_threshold': 0.5400954892324498]],
        'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 92.
→70022357199642,
        'MCC Binary': inf,
        'Gini': inf,
        'AUC': 27.249798224374487,
        'Precision': 88.88888888888889,
        'F1': 40.26936026936027,
        'Balanced Accuracy Binary': 12.812752219531875,
        'Accuracy Binary': 2.5026383235338567},
        'percent_better_than_baseline': 92.70022357199642,
        'validation_score': 0.27690265474464915},
4: {'id': 4,
    'pipeline_name': 'LightGBM Classifier w/ Drop Columns Transformer + Imputer +
→DateTime Featurization Component + One Hot Encoder + Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
→BinaryClassificationPipeline,
    'pipeline_summary': 'LightGBM Classifier w/ Drop Columns Transformer + Imputer +
→DateTime Featurization Component + One Hot Encoder + Oversampler',
    'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
        'Imputer': {'categorical_impute_strategy': 'most_frequent',
            'numeric_impute_strategy': 'mean',
            'categorical_fill_value': None,
            'numeric_fill_value': None},
        'DateTime Featurization Component': {'features_to_extract': ['year',
            'month',
            'day_of_week',
            'hour'],
            'encode_as_categories': False,
            'date_index': None},
        'One Hot Encoder': {'top_n': 10,
            'features_to_encode': None,
            'categories': None,
            'drop': 'if_binary',
            'handle_unknown': 'ignore',
            'handle_missing': 'error'},
        'Oversampler': {'sampling_ratio': 0.25,
            'k_neighbors_default': 5,
            'n_jobs': -1,
            'sampling_ratio_dict': None,
            'k_neighbors': 5},
        'LightGBM Classifier': {'boosting_type': 'gbdt',
            'learning_rate': 0.1,
            'n_estimators': 100,
            'max_depth': 0,
            'num_leaves': 31,
            'min_child_samples': 20,
            'n_jobs': -1,
            'bagging_freq': 0,
            'bagging_fraction': 0.9}},
    'mean_cv_score': 0.3208641239535214,
    'standard_deviation_cv_score': 0.14757715366385865,
    'high_variance_cv': False,
    'training_time': 1.6711151599884033,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',

```

(continues on next page)

(continued from previous page)

```

        0.22676504669023242),
        ('MCC Binary', 0.3342789617107607),
        ('Gini', 0.771186440677966),
        ('AUC', 0.885593220338983),
        ('Precision', 1.0),
        ('F1', 0.2222222222222222),
        ('Balanced Accuracy Binary', 0.5625),
        ('Accuracy Binary', 0.8955223880597015),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.22676504669023242,
    'binary_classification_threshold': 0.8315193475327842},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
        0.49095040111140575),
        ('MCC Binary', 0.36540212737375455),
        ('Gini', 0.4576271186440677),
        ('AUC', 0.7288135593220338),
        ('Precision', 0.6666666666666666),
        ('F1', 0.36363636363636365),
        ('Balanced Accuracy Binary', 0.6165254237288136),
        ('Accuracy Binary', 0.8955223880597015),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.49095040111140575,
    'binary_classification_threshold': 0.812705509811428},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
        0.24487692405892594),
        ('MCC Binary', 0.3600976668493281),
        ('Gini', 0.7627118644067796),
        ('AUC', 0.8813559322033898),
        ('Precision', 1.0),
        ('F1', 0.25),
        ('Balanced Accuracy Binary', 0.5714285714285714),
        ('Accuracy Binary', 0.9090909090909091),
        ('# Training', 134),
        ('# Validation', 66)]),
    'mean_cv_score': 0.24487692405892594,
    'binary_classification_threshold': 0.9574727501533578}],
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 91.
→9186416958576,
    'MCC Binary': inf,
    'Gini': inf,
    'AUC': 33.19209039548022,
    'Precision': 88.88888888888889,
    'F1': 27.86195286195286,
    'Balanced Accuracy Binary': 8.348466505246165,
    'Accuracy Binary': 1.5000753806724054},
'percent_better_than_baseline': 91.9186416958576,
'validation_score': 0.22676504669023242},
5: {'id': 5,
    'pipeline_name': 'Logistic Regression Classifier w/ Drop Columns Transformer +
→Imputer + DateTime Featurization Component + One Hot Encoder + Oversampler +
→Standard Scaler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
→BinaryClassificationPipeline,
    'pipeline_summary': 'Logistic Regression Classifier w/ Drop Columns Transformer +
→Imputer + DateTime Featurization Component + One Hot Encoder + Oversampler +
→Standard Scaler',

```

(continues on next page)

(continued from previous page)

```

'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
'Imputer': {'categorical_impute_strategy': 'most_frequent',
'numeric_impute_strategy': 'mean',
'categorical_fill_value': None,
'numeric_fill_value': None},
'DateTime Featurization Component': {'features_to_extract': ['year',
'month',
'day_of_week',
'hour']},
'encode_as_categories': False,
'date_index': None},
'One Hot Encoder': {'top_n': 10,
'features_to_encode': None,
'categories': None,
'drop': 'if_binary',
'handle_unknown': 'ignore',
'handle_missing': 'error'},
'Oversampler': {'sampling_ratio': 0.25,
'k_neighbors_default': 5,
'n_jobs': -1,
'sampling_ratio_dict': None,
'k_neighbors': 5},
'Logistic Regression Classifier': {'penalty': 'l2',
'C': 1.0,
'n_jobs': -1,
'multi_class': 'auto',
'solver': 'lbfgs'}},
'mean_cv_score': 0.6516020967013594,
'standard_deviation_cv_score': 0.11741047129130076,
'high_variance_cv': False,
'training_time': 3.209779739379883,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.5224083958088352),
('MCC Binary', 0.045713245813732276),
('Gini', 0.15254237288135597),
('AUC', 0.576271186440678),
('Precision', 0.16666666666666666),
('F1', 0.14285714285714288),
('Balanced Accuracy Binary', 0.5201271186440678),
('Accuracy Binary', 0.8208955223880597),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.5224083958088352,
'binary_classification_threshold': 0.36567760969764035},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.6806045725061826),
('MCC Binary', 0.17518582316850065),
('Gini', 0.15677966101694918),
('AUC', 0.5783898305084746),
('Precision', 0.2857142857142857),
('F1', 0.26666666666666666),
('Balanced Accuracy Binary', 0.5826271186440678),
('Accuracy Binary', 0.835820895522388),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.6806045725061826,
'binary_classification_threshold': 0.21868425771854014},

```

(continues on next page)

(continued from previous page)

```

{'all_objective_scores': OrderedDict([('Log Loss Binary',
                                     0.7517933217890604),
                                     ('MCC Binary', 0.04803502496528411),
                                     ('Gini', 0.07021791767554486),
                                     ('AUC', 0.5351089588377724),
                                     ('Precision', 0.13333333333333333),
                                     ('F1', 0.18181818181818182),
                                     ('Balanced Accuracy Binary', 0.5326876513317191),
                                     ('Accuracy Binary', 0.7272727272727273),
                                     ('# Training', 134),
                                     ('# Validation', 66)]),
 'mean_cv_score': 0.7517933217890604,
 'binary_classification_threshold': 0.44118090538932764}},
 'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 83.
↪58859834408628,
 'MCC Binary': inf,
 'Gini': inf,
 'AUC': 6.325665859564166,
 'Precision': 19.523809523809522,
 'F1': 19.71139971139971,
 'Balanced Accuracy Binary': 4.514729620661829,
 'Accuracy Binary': -9.038142620232147},
 'percent_better_than_baseline': 83.58859834408628,
 'validation_score': 0.5224083958088352},
 6: {'id': 6,
     'pipeline_name': 'XGBoost Classifier w/ Drop Columns Transformer + Imputer + ↪
↪DateTime Featurization Component + One Hot Encoder + Oversampler',
     'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
     'pipeline_summary': 'XGBoost Classifier w/ Drop Columns Transformer + Imputer + ↪
↪DateTime Featurization Component + One Hot Encoder + Oversampler',
     'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
                    'Imputer': {'categorical_impute_strategy': 'most_frequent',
                                'numeric_impute_strategy': 'mean',
                                'categorical_fill_value': None,
                                'numeric_fill_value': None},
                    'DateTime Featurization Component': {'features_to_extract': ['year',
                                        'month',
                                        'day_of_week',
                                        'hour'],
                    'encode_as_categories': False,
                    'date_index': None},
                    'One Hot Encoder': {'top_n': 10,
                                        'features_to_encode': None,
                                        'categories': None,
                                        'drop': 'if_binary',
                                        'handle_unknown': 'ignore',
                                        'handle_missing': 'error'},
                    'Oversampler': {'sampling_ratio': 0.25,
                                    'k_neighbors_default': 5,
                                    'n_jobs': -1,
                                    'sampling_ratio_dict': None,
                                    'k_neighbors': 5},
                    'XGBoost Classifier': {'eta': 0.1,
                                        'max_depth': 6,
                                        'min_child_weight': 1,
                                        'n_estimators': 100,

```

(continues on next page)

(continued from previous page)

```

    'n_jobs': -1,
    'eval_metric': 'logloss'}}},
'mean_cv_score': 0.27376580269962625,
'standard_deviation_cv_score': 0.16442482847286716,
'high_variance_cv': False,
'training_time': 2.0751006603240967,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.20639750014922215),
    ('MCC Binary', 0.0),
    ('Gini', 0.7118644067796611),
    ('AUC', 0.8559322033898306),
    ('Precision', 0.0),
    ('F1', 0.0),
    ('Balanced Accuracy Binary', 0.5),
    ('Accuracy Binary', 0.8805970149253731),
    ('# Training', 133),
    ('# Validation', 67)]),
'mean_cv_score': 0.20639750014922215,
'binary_classification_threshold': 0.9446923046727642},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.46117587266435733),
    ('MCC Binary', 0.36540212737375455),
    ('Gini', 0.40042372881355925),
    ('AUC', 0.7002118644067796),
    ('Precision', 0.6666666666666666),
    ('F1', 0.36363636363636365),
    ('Balanced Accuracy Binary', 0.6165254237288136),
    ('Accuracy Binary', 0.8955223880597015),
    ('# Training', 133),
    ('# Validation', 67)]),
'mean_cv_score': 0.46117587266435733,
'binary_classification_threshold': 0.8627784224489395},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.15372403528529918),
    ('MCC Binary', 0.3600976668493281),
    ('Gini', 0.8111380145278451),
    ('AUC', 0.9055690072639225),
    ('Precision', 1.0),
    ('F1', 0.25),
    ('Balanced Accuracy Binary', 0.5714285714285714),
    ('Accuracy Binary', 0.9090909090909091),
    ('# Training', 134),
    ('# Validation', 66)]),
'mean_cv_score': 0.15372403528529918,
'binary_classification_threshold': 0.9306861087519589}},
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 93.
↪10487094731317,
'MCC Binary': inf,
'Gini': inf,
'AUC': 32.05710250201776,
'Precision': 55.55555555555555,
'F1': 20.454545454545457,
'Balanced Accuracy Binary': 6.265133171912829,
'Accuracy Binary': 1.0025629428614624},
'percent_better_than_baseline': 93.10487094731317,
'validation_score': 0.20639750014922215},
7: {'id': 7,

```

(continues on next page)

(continued from previous page)

```

'pipeline_name': 'Extra Trees Classifier w/ Drop Columns Transformer + Imputer +
↳ DateTime Featurization Component + One Hot Encoder + Oversampler',
'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↳ BinaryClassificationPipeline,
'pipeline_summary': 'Extra Trees Classifier w/ Drop Columns Transformer + Imputer
↳ + DateTime Featurization Component + One Hot Encoder + Oversampler',
'parameters': {'Drop Columns Transformer': {'columns': ['currency']}},
'Imputer': {'categorical_impute_strategy': 'most_frequent',
'numeric_impute_strategy': 'mean',
'categorical_fill_value': None,
'numeric_fill_value': None},
'Datetime Featurization Component': {'features_to_extract': ['year',
'month',
'day_of_week',
'hour'],
'encode_as_categories': False,
'date_index': None},
'One Hot Encoder': {'top_n': 10,
'features_to_encode': None,
'categories': None,
'drop': 'if_binary',
'handle_unknown': 'ignore',
'handle_missing': 'error'},
'Oversampler': {'sampling_ratio': 0.25,
'k_neighbors_default': 5,
'n_jobs': -1,
'sampling_ratio_dict': None,
'k_neighbors': 5},
'Extra Trees Classifier': {'n_estimators': 100,
'max_features': 'auto',
'max_depth': 6,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_jobs': -1}},
'mean_cv_score': 0.3386759195464455,
'standard_deviation_cv_score': 0.030914031011629657,
'high_variance_cv': False,
'training_time': 1.9855828285217285,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.3144942839064824),
('MCC Binary', 0.16779531085172097),
('Gini', 0.4872881355932204),
('AUC', 0.7436440677966102),
('Precision', 0.17142857142857143),
('F1', 0.2790697674418604),
('Balanced Accuracy Binary', 0.6292372881355932),
('Accuracy Binary', 0.5373134328358209),
('# Training', 133),
('# Validation', 67)])},
'mean_cv_score': 0.3144942839064824,
'binary_classification_threshold': 0.09086573203500445},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.3735068711989279),
('MCC Binary', -0.0645917880798583),
('Gini', 0.19491525423728828),
('AUC', 0.5974576271186441),
('Precision', 0.0),

```

(continues on next page)

(continued from previous page)

```

        ('F1', 0.0),
        ('Balanced Accuracy Binary', 0.4830508474576271),
        ('Accuracy Binary', 0.8507462686567164),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.3735068711989279,
    'binary_classification_threshold': 0.2616991380282906},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.3280266035339262),
    ('MCC Binary', 0.2005853178066611),
    ('Gini', 0.35593220338983045),
    ('AUC', 0.6779661016949152),
    ('Precision', 0.23076923076923078),
    ('F1', 0.3),
    ('Balanced Accuracy Binary', 0.6295399515738499),
    ('Accuracy Binary', 0.7878787878787878),
    ('# Training', 134),
    ('# Validation', 66)]),
    'mean_cv_score': 0.3280266035339262,
    'binary_classification_threshold': 0.22061716793337022}},
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 91.
↪47002967762083,
    'MCC Binary': inf,
    'Gini': inf,
    'AUC': 17.302259887005654,
    'Precision': 13.406593406593407,
    'F1': 19.302325581395344,
    'Balanced Accuracy Binary': 8.06093623890235,
    'Accuracy Binary': -15.973164480627155},
'percent_better_than_baseline': 91.47002967762083,
'validation_score': 0.3144942839064824},
8: {'id': 8,
    'pipeline_name': 'CatBoost Classifier w/ Drop Columns Transformer + Imputer +
↪DateTime Featurization Component + Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'CatBoost Classifier w/ Drop Columns Transformer + Imputer +
↪DateTime Featurization Component + Oversampler',
    'parameters': {'Drop Columns Transformer': {'columns': ['currency']},
    'Imputer': {'categorical_impute_strategy': 'most_frequent',
    'numeric_impute_strategy': 'mean',
    'categorical_fill_value': None,
    'numeric_fill_value': None},
    'DateTime Featurization Component': {'features_to_extract': ['year',
    'month',
    'day_of_week',
    'hour'],
    'encode_as_categories': False,
    'date_index': None},
    'Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'categorical_features': [3, 4, 7, 8],
    'k_neighbors': 5},
    'CatBoost Classifier': {'n_estimators': 10,
    'eta': 0.03,

```

(continues on next page)

(continued from previous page)

```

    'max_depth': 6,
    'bootstrap_type': None,
    'silent': True,
    'allow_writing_files': False,
    'n_jobs': -1}},
'mean_cv_score': 0.6269683475365982,
'standard_deviation_cv_score': 0.005110376657691124,
'high_variance_cv': False,
'training_time': 1.126361608505249,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.6328217109569406),
    ('MCC Binary', 0.7712055916006633),
    ('Gini', 0.7669491525423728),
    ('AUC', 0.8834745762711864),
    ('Precision', 1.0),
    ('F1', 0.7692307692307693),
    ('Balanced Accuracy Binary', 0.8125),
    ('Accuracy Binary', 0.9552238805970149),
    ('# Training', 133),
    ('# Validation', 67)]),
'mean_cv_score': 0.6328217109569406,
'binary_classification_threshold': 0.49572877036722973},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.6246893942305619),
    ('MCC Binary', 0.47636443708895493),
    ('Gini', 0.31779661016949157),
    ('AUC', 0.6588983050847458),
    ('Precision', 1.0),
    ('F1', 0.4),
    ('Balanced Accuracy Binary', 0.625),
    ('Accuracy Binary', 0.9104477611940298),
    ('# Training', 133),
    ('# Validation', 67)]),
'mean_cv_score': 0.6246893942305619,
'binary_classification_threshold': 0.5136041807822178},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.6233939374222923),
    ('MCC Binary', 0.5132181379714255),
    ('Gini', 0.937046004842615),
    ('AUC', 0.9685230024213075),
    ('Precision', 1.0),
    ('F1', 0.4444444444444445),
    ('Balanced Accuracy Binary', 0.6428571428571428),
    ('Accuracy Binary', 0.9242424242424242),
    ('# Training', 134),
    ('# Validation', 66)]),
'mean_cv_score': 0.6233939374222923,
'binary_classification_threshold': 0.4978114819159809}],
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 84.
→20902966847964,
'MCC Binary': inf,
'Gini': inf,
'AUC': 33.69652945924132,
'Precision': 100.0,
'F1': 53.789173789173795,
'Balanced Accuracy Binary': 19.345238095238095,
'Accuracy Binary': 4.492688074777629},

```

(continues on next page)

(continued from previous page)

```
'percent_better_than_baseline': 84.20902966847964,
'validation_score': 0.6328217109569406}},
'search_order': [0, 1, 2, 3, 4, 5, 6, 7, 8]}
```

4.1.12 Parallel AutoML

By default, all pipelines in an AutoML batch are evaluated in series. Pipelines can be evaluated in parallel to improve performance during AutoML search. This is accomplished by a futures style submission and evaluation of pipelines in a batch. As of this writing, the pipelines use a threaded model for concurrent evaluation. This is similar to the currently implemented `n_jobs` parameter in the estimators, which uses increased numbers of threads to train and evaluate estimators.

Quick Start

To quickly use some parallelism to enhance the pipeline searching, a string can be passed through to `AutoMLSearch` during initialization to setup the parallel engine and client within the `AutoMLSearch` object. The current options are “`cf_threaded`”, “`cf_process`”, “`dask_threaded`” and “`dask_process`” and indicate the futures backend to use and whether to use threaded- or process-level parallelism.

```
[26]: automl_cf_threaded = AutoMLSearch(X_train=X, y_train=y,
                                     problem_type="binary",
                                     allowed_model_families=[ModelFamily.LINEAR_MODEL],
                                     engine="cf_threaded")
automl_cf_threaded.search(show_iteration_plot = False)
automl_cf_threaded.close_engine()
```

Parallelism with Concurrent Futures

The `EngineBase` class is robust and extensible enough to support futures-like implementations from a variety of libraries. The `CFEngine` extends the `EngineBase` to use the native Python [concurrent.futures library](#). The `CFEngine` supports both thread- and process-level parallelism. The type of parallelism can be chosen using either the `ThreadPoolExecutor` or the `ProcessPoolExecutor`. If either executor is passed a `max_workers` parameter, it will set the number of processes and threads spawned. If not, the default number of processes will be equal to the number of processors available and the number of threads set to five times the number of processors available.

Here, the `CFEngine` is invoked with default parameters, which is threaded parallelism using all available threads.

```
[27]: from concurrent.futures import ThreadPoolExecutor

from evalml.automl.engine.cf_engine import CFEngine, CFClient

cf_engine = CFEngine(CFClient(ThreadPoolExecutor(max_workers=4)))
automl_cf_threaded = AutoMLSearch(X_train=X, y_train=y,
                                problem_type="binary",
                                allowed_model_families=[ModelFamily.LINEAR_MODEL],
                                engine=cf_engine)
automl_cf_threaded.search(show_iteration_plot = False)
automl_cf_threaded.close_engine()
```

Note: the cell demonstrating process-level parallelism is a markdown due to incompatibility with our ReadTheDocs build. It can be run successfully locally.

```

from concurrent.futures import ProcessPoolExecutor

# Repeat the process but using process-level parallelism\
cf_engine = CFEngine(CFClient(ProcessPoolExecutor(max_workers=2)))
automl_cf_process = AutoMLSearch(X_train=X, y_train=y,
                                problem_type="binary",
                                engine="cf_process")
automl_cf_process.search(show_iteration_plot = False)
automl_cf_process.close_engine()

```

Parallelism with Dask

Thread or process level parallelism can be explicitly invoked for the DaskEngine (as well as the CFEngine). The processes can be set to True and the number of processes set using n_workers. If processes is set to False, then the resulting parallelism will be threaded and n_workers will represent the threads used. Examples of both follow.

```

[28]: from dask.distributed import LocalCluster

from evalml.automl.engine import DaskEngine

dask_engine_p2 = DaskEngine(cluster=LocalCluster(processes=True, n_workers = 2))
automl_dask_p2 = AutoMLSearch(X_train=X, y_train=y,
                              problem_type="binary",
                              allowed_model_families=[ModelFamily.LINEAR_MODEL],
                              engine=dask_engine_p2)
automl_dask_p2.search(show_iteration_plot = False)

# Explicitly shutdown the automl object's LocalCluster
automl_dask_p2.close_engine()

Loky-backed parallel loops cannot be called in a multiprocessing, setting n_jobs=1
Loky-backed parallel loops cannot be called in a multiprocessing, setting n_jobs=1
Loky-backed parallel loops cannot be called in a multiprocessing, setting n_jobs=1
Loky-backed parallel loops cannot be called in a multiprocessing, setting n_jobs=1

```

```

[29]: dask_engine_t4 = DaskEngine(cluster=LocalCluster(processes=False, n_workers = 4))

automl_dask_t4 = AutoMLSearch(X_train=X, y_train=y,
                              problem_type="binary",
                              allowed_model_families=[ModelFamily.LINEAR_MODEL],
                              engine=dask_engine_t4)
automl_dask_t4.search(show_iteration_plot = False)
automl_dask_t4.close_engine()

```

As we can see, a significant performance gain can result from simply using something other than the default SequentialEngine, ranging from a 100% speed up with multiple processes to 500% speedup with multiple threads!

```

[30]: print("Sequential search duration: %s" % str(automl.search_duration))
print("Concurrent futures (threaded) search duration: %s" % str(automl_cf_threaded.
    ↳ search_duration))
print("Dask (two processes) search duration: %s" % str(automl_dask_p2.search_
    ↳ duration))
print("Dask (four threads) search duration: %s" % str(automl_dask_t4.search_duration))

```

```
Sequential search duration: 17.823562383651733
Concurrent futures (threaded) search duration: 4.28137731552124
Dask (two processes) search duration: 9.347124814987183
Dask (four threads) search duration: 4.964043617248535
```

4.2 Pipelines

EvalML pipelines represent a sequence of operations to be applied to data, where each operation is either a data transformation or an ML modeling algorithm.

A pipeline holds a combination of one or more components, which will be applied to new input data in sequence.

Each component and pipeline supports a set of parameters which configure its behavior. The AutoML search process seeks to find the combination of pipeline structure and pipeline parameters which perform the best on the data.

4.2.1 Defining a Pipeline Instance

Pipeline instances can be instantiated using any of the following classes:

- `RegressionPipeline`
- `BinaryClassificationPipeline`
- `MulticlassClassificationPipeline`
- `TimeSeriesRegressionPipeline`
- `TimeSeriesBinaryClassificationPipeline`
- `TimeSeriesMulticlassClassificationPipeline`

The class you want to use will depend on your problem type. The only required parameter input for instantiating a pipeline instance is `component_graph`, which is either a list or a dictionary containing a sequence of components to be fit and evaluated.

A `component_graph` list is the default representation, which represents a linear order of transforming components with an estimator as the final component. A `component_graph` dictionary is used to represent a non-linear graph of components, where the key is a unique name for each component and the value is a list with the component's class as the first element and any parents of the component as the following element(s). For either `component_graph` format, each component can be provided as a reference to the component class for custom components, and as either a string name or as a reference to the component class for components defined in EvalML.

```
[1]: from evalml.pipelines import MulticlassClassificationPipeline

component_graph_as_list = ['Imputer', 'Random Forest Classifier']
MulticlassClassificationPipeline(component_graph=component_graph_as_list)

[1]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↪', 'y'], 'Random Forest Classifier': ['Random Forest Classifier', 'Imputer.x', 'y']})
↪, parameters={'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_
↪impute_strategy': 'mean', 'categorical_fill_value': None, 'numeric_fill_value': _
↪None}, 'Random Forest Classifier':{'n_estimators': 100, 'max_depth': 6, 'n_jobs': -
↪1}}, random_seed=0)
```



```
[2]: component_graph_as_dict = {
    'Imputer': ['Imputer', 'X', 'y'],
    'Encoder': ['One Hot Encoder', 'Imputer.x', 'y'],
    'Random Forest Clf': ['Random Forest Classifier', 'Encoder.x', 'y'],
    'Elastic Net Clf': ['Elastic Net Classifier', 'Encoder.x', 'y'],
    'Final Estimator': ['Logistic Regression Classifier', 'Random Forest Clf.x',
↳ 'Elastic Net Clf.x', 'y']
}

MulticlassClassificationPipeline(component_graph=component_graph_as_dict)

[2]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↳ ', 'y'], 'Encoder': ['One Hot Encoder', 'Imputer.x', 'y'], 'Random Forest Clf': [
↳ 'Random Forest Classifier', 'Encoder.x', 'y'], 'Elastic Net Clf': ['Elastic Net_
↳ Classifier', 'Encoder.x', 'y'], 'Final Estimator': ['Logistic Regression Classifier
↳ ', 'Random Forest Clf.x', 'Elastic Net Clf.x', 'y']}, parameters={'Imputer':{
↳ 'categorical_impute_strategy': 'most_frequent', 'numeric_impute_strategy': 'mean',
↳ 'categorical_fill_value': None, 'numeric_fill_value': None}, 'Encoder':{'top_n': 10,
↳ 'features_to_encode': None, 'categories': None, 'drop': 'if_binary', 'handle_
↳ unknown': 'ignore', 'handle_missing': 'error'}, 'Random Forest Clf':{'n_estimators':
↳ 100, 'max_depth': 6, 'n_jobs': -1}, 'Elastic Net Clf':{'penalty': 'elasticnet', 'C
↳ ': 1.0, 'l1_ratio': 0.15, 'n_jobs': -1, 'multi_class': 'auto', 'solver': 'saga'},
↳ 'Final Estimator':{'penalty': 'l2', 'C': 1.0, 'n_jobs': -1, 'multi_class': 'auto',
↳ 'solver': 'lbfgs'}}}, random_seed=0)
```

If you're using your own *custom components* you can refer to them like so:

```
[3]: from evalml.pipelines.components import Transformer

class NewTransformer(Transformer):
    name = 'New Transformer'
    hyperparameter_ranges = {
        "parameter_1": ['a', 'b', 'c']
    }

    def __init__(self, parameter_1=1, random_seed=0):
        parameters = {"parameter_1": parameter_1}
        super().__init__(parameters=parameters,
            random_seed=random_seed)

MulticlassClassificationPipeline([NewTransformer, 'Random Forest Classifier'])

[3]: pipeline = MulticlassClassificationPipeline(component_graph={'New Transformer':_
↳ [NewTransformer, 'X', 'y'], 'Random Forest Classifier': ['Random Forest Classifier',
↳ 'New Transformer.x', 'y']}, parameters={'New Transformer':{'parameter_1': 1},
↳ 'Random Forest Classifier':{'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}},_
↳ random_seed=0)
```

4.2.2 Pipeline Usage

All pipelines define the following methods:

- `fit` fits each component on the provided training data, in order.
- `predict` computes the predictions of the component graph on the provided data.
- `score` computes the value of *an objective* on the provided data.

```
[4]: from evalml.demos import load_wine
X, y = load_wine()

pipeline = MulticlassClassificationPipeline(['Imputer', 'Random Forest Classifier'])
pipeline.fit(X, y)
print(pipeline.predict(X))
print(pipeline.score(X, y, objectives=['log loss multiclass']))
```

```

      Number of Features
Numeric                13

Number of training examples: 178
Targets
class_1    39.89%
class_0    33.15%
class_2    26.97%
Name: target, dtype: object
0      class_0
1      class_0
2      class_0
3      class_0
4      class_0
...
173     class_2
174     class_2
175     class_2
176     class_2
177     class_2
Length: 178, dtype: category
Categories (3, object): ['class_0', 'class_1', 'class_2']
OrderedDict([('Log Loss Multiclass', 0.04132737017536148)])
```

4.2.3 Custom Name

By default, a pipeline's name is created using the component graph that makes up the pipeline. E.g. A pipeline with an imputer, one-hot encoder, and logistic regression classifier will have the name 'Logistic Regression Classifier w/ Imputer + One Hot Encoder'.

If you'd like to override the pipeline's name attribute, you can set the `custom_name` parameter when initializing a pipeline, like so:

```
[5]: component_graph = ['Imputer', 'One Hot Encoder', 'Logistic Regression Classifier']
pipeline = MulticlassClassificationPipeline(component_graph)
print("Pipeline with default name:", pipeline.name)
```

(continues on next page)

(continued from previous page)

```

pipeline_with_name = MulticlassClassificationPipeline(component_graph, custom_name=
↳ "My cool custom pipeline")
print("Pipeline with custom name:", pipeline_with_name.name)

Pipeline with default name: Logistic Regression Classifier w/ Imputer + One Hot_
↳ Encoder
Pipeline with custom name: My cool custom pipeline

```

4.2.4 Pipeline Parameters

You can also pass in custom parameters by using the `parameters` parameter, which will then be used when instantiating each component in `component_graph`. The parameters dictionary needs to be in the format of a two-layered dictionary where the key-value pairs are the component name and corresponding component parameters dictionary. The component parameters dictionary consists of (parameter name, parameter values) key-value pairs.

An example will be shown below. The API reference for component parameters can also be found [\[here\]](#) (`../api_reference.rst#components`).

```

[6]: parameters = {
    'Imputer': {
        "categorical_impute_strategy": "most_frequent",
        "numeric_impute_strategy": "median"
    },
    'Logistic Regression Classifier': {
        'penalty': 'l2',
        'C': 1.0,
    }
}

component_graph = ['Imputer', 'One Hot Encoder', 'Standard Scaler', 'Logistic_
↳ Regression Classifier']
MulticlassClassificationPipeline(component_graph=component_graph,
↳ parameters=parameters)

[6]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↳ ', 'y'], 'One Hot Encoder': ['One Hot Encoder', 'Imputer.x', 'y'], 'Standard Scaler
↳ ': ['Standard Scaler', 'One Hot Encoder.x', 'y'], 'Logistic Regression Classifier':
↳ ['Logistic Regression Classifier', 'Standard Scaler.x', 'y']}, parameters={'Imputer
↳ ': {'categorical_impute_strategy': 'most_frequent', 'numeric_impute_strategy':
↳ 'median', 'categorical_fill_value': None, 'numeric_fill_value': None}, 'One Hot_
↳ Encoder': {'top_n': 10, 'features_to_encode': None, 'categories': None, 'drop': 'if_
↳ binary', 'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'Logistic_
↳ Regression Classifier': {'penalty': 'l2', 'C': 1.0, 'n_jobs': -1, 'multi_class':
↳ 'auto', 'solver': 'lbfgs'}}}, random_seed=0)

```

4.2.5 Pipeline Description

You can call `.graph()` to see each component and its parameters. Each component takes in data and feeds it to the next.

```

[7]: component_graph = ['Imputer', 'One Hot Encoder', 'Standard Scaler', 'Logistic_
↳ Regression Classifier']
pipeline = MulticlassClassificationPipeline(component_graph=component_graph,
↳ parameters=parameters)
pipeline.graph()

```

```
[7]:
[8]: component_graph_as_dict = {
    'Imputer': ['Imputer', 'X', 'y'],
    'Encoder': ['One Hot Encoder', 'Imputer.x', 'y'],
    'Random Forest Clf': ['Random Forest Classifier', 'Encoder.x', 'y'],
    'Elastic Net Clf': ['Elastic Net Classifier', 'Encoder.x', 'y'],
    'Final Estimator': ['Logistic Regression Classifier', 'Random Forest Clf.x',
↳ 'Elastic Net Clf.x', 'y']
}

nonlinear_pipeline = MulticlassClassificationPipeline(component_graph=component_graph_
↳ as_dict)
nonlinear_pipeline.graph()
```

[8]: You can see a textual representation of the pipeline by calling `.describe()`:

```
[9]: pipeline.describe()

*****
* Logistic Regression Classifier w/ Imputer + One Hot Encoder + Standard Scaler *
*****

Problem Type: multiclass
Model Family: Linear

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : median
    * categorical_fill_value : None
    * numeric_fill_value : None
2. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
3. Standard Scaler
4. Logistic Regression Classifier
    * penalty : l2
    * C : 1.0
    * n_jobs : -1
    * multi_class : auto
    * solver : lbfgs
```

```
[10]: nonlinear_pipeline.describe()

*****
* Logistic Regression Classifier w/ Imputer + One Hot Encoder + Random Forest_
↳ Classifier + Elastic Net Classifier *
*****
```

(continues on next page)

(continued from previous page)

```

Problem Type: multiclass
Model Family: Linear

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
3. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6
    * n_jobs : -1
4. Elastic Net Classifier
    * penalty : elasticnet
    * C : 1.0
    * l1_ratio : 0.15
    * n_jobs : -1
    * multi_class : auto
    * solver : saga
5. Logistic Regression Classifier
    * penalty : l2
    * C : 1.0
    * n_jobs : -1
    * multi_class : auto
    * solver : lbfgs

```

4.2.6 Component Graph

You can use `pipeline.get_component(name)` and provide the component name to access any component (API reference [here](#)):

```

[11]: pipeline.get_component('Imputer')
[11]: Imputer(categorical_impute_strategy='most_frequent', numeric_impute_strategy='median',
    ↪ categorical_fill_value=None, numeric_fill_value=None)

[12]: nonlinear_pipeline.get_component('Elastic Net Clf')
[12]: ElasticNetClassifier(penalty='elasticnet', C=1.0, l1_ratio=0.15, n_jobs=-1, multi_
    ↪ class='auto', solver='saga')

```

Alternatively, you can index directly into the pipeline to get a component

```

[13]: first_component = pipeline[0]
    print(first_component.name)

```

```
Imputer
```

```
[14]: nonlinear_pipeline['Final Estimator']
```

```
[14]: LogisticRegressionClassifier(penalty='l2', C=1.0, n_jobs=-1, multi_class='auto',  
↳ solver='lbfgs')
```

4.2.7 Pipeline Estimator

EvalML enforces that the last component of a linear pipeline is an estimator. You can access this estimator directly by using `pipeline.estimator`.

```
[15]: pipeline.estimator
```

```
[15]: LogisticRegressionClassifier(penalty='l2', C=1.0, n_jobs=-1, multi_class='auto',  
↳ solver='lbfgs')
```

4.2.8 Input Feature Names

After a pipeline is fitted, you can access a pipeline's `input_feature_names` attribute to obtain a dictionary containing a list of feature names passed to each component of the pipeline. This could be especially useful for debugging where a feature might have been dropped or detecting unexpected behavior.

```
[16]: pipeline = MulticlassClassificationPipeline(['Imputer', 'Random Forest Classifier'])  
pipeline.fit(X, y)  
pipeline.input_feature_names
```

```
[16]: {'Imputer': ['alcohol',  
    'malic_acid',  
    'ash',  
    'alcalinity_of_ash',  
    'magnesium',  
    'total_phenols',  
    'flavanoids',  
    'nonflavanoid_phenols',  
    'proanthocyanins',  
    'color_intensity',  
    'hue',  
    'od280/od315_of_diluted_wines',  
    'proline'],  
    'Random Forest Classifier': ['alcohol',  
    'malic_acid',  
    'ash',  
    'alcalinity_of_ash',  
    'magnesium',  
    'total_phenols',  
    'flavanoids',  
    'nonflavanoid_phenols',  
    'proanthocyanins',  
    'color_intensity',  
    'hue',  
    'od280/od315_of_diluted_wines',  
    'proline']}
```

4.2.9 Saving and Loading Pipelines

You can save and load trained or untrained pipeline instances using the Python `pickle` format, like so:

```
[17]: import pickle

pipeline_to_pickle = MulticlassClassificationPipeline(['Imputer', 'Random Forest_
↳Classifier'])

with open("pipeline.pkl", 'wb') as f:
    pickle.dump(pipeline_to_pickle, f)

pickled_pipeline = None
with open('pipeline.pkl', 'rb') as f:
    pickled_pipeline = pickle.load(f)

assert pickled_pipeline == pipeline_to_pickle
pickled_pipeline.fit(X, y)

[17]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↳', 'y'], 'Random Forest Classifier': ['Random Forest Classifier', 'Imputer.x', 'y']}
↳, parameters={'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_
↳impute_strategy': 'mean', 'categorical_fill_value': None, 'numeric_fill_value':_
↳None}, 'Random Forest Classifier':{'n_estimators': 100, 'max_depth': 6, 'n_jobs': -
↳1}}, random_seed=0)
```

4.2.10 Generate Code

Once you have instantiated a pipeline, you can generate string Python code to recreate this pipeline, which can then be saved and run elsewhere with EvalML. `generate_pipeline_code` requires a pipeline instance as the input. It can also handle custom components, but it won't return the code required to define the component. Note that any external libraries used in creating the pipeline instance will also need to be imported to execute the returned code.

Code generation is not yet supported for nonlinear pipelines.

```
[18]: from evalml.pipelines.utils import generate_pipeline_code
from evalml.pipelines import MulticlassClassificationPipeline
import pandas as pd
from evalml.utils import infer_feature_types
from skopt.space import Integer

class MyDropNullColumns(Transformer):
    """Transformer to drop features whose percentage of NaN values exceeds a_
↳specified threshold"""
    name = "My Drop Null Columns Transformer"
    hyperparameter_ranges = {}

    def __init__(self, pct_null_threshold=1.0, random_seed=0, **kwargs):
        """Initializes an transformer to drop features whose percentage of NaN values_
↳exceeds a specified threshold.

        Arguments:
            pct_null_threshold(float): The percentage of NaN values in an input_
↳feature to drop.
            Must be a value between [0, 1] inclusive. If equal to 0.0, will drop_
↳columns with any null values.
            If equal to 1.0, will drop columns with all null values. Defaults to_
↳0.95.
        """
```

(continues on next page)

(continued from previous page)

```

    """
    if pct_null_threshold < 0 or pct_null_threshold > 1:
        raise ValueError("pct_null_threshold must be a float between 0 and 1,
↪inclusive.")
    parameters = {"pct_null_threshold": pct_null_threshold}
    parameters.update(kwargs)

    self._cols_to_drop = None
    super().__init__(parameters=parameters,
                     component_obj=None,
                     random_seed=random_seed)

    def fit(self, X, y=None):
        pct_null_threshold = self.parameters["pct_null_threshold"]
        X = infer_feature_types(X)
        percent_null = X.isnull().mean()
        if pct_null_threshold == 0.0:
            null_cols = percent_null[percent_null > 0]
        else:
            null_cols = percent_null[percent_null >= pct_null_threshold]
        self._cols_to_drop = list(null_cols.index)
        return self

    def transform(self, X, y=None):
        """Transforms data X by dropping columns that exceed the threshold of null_
↪values.
        Arguments:
            X (pd.DataFrame): Data to transform
            y (pd.Series, optional): Targets
        Returns:
            pd.DataFrame: Transformed X
        """

        X = infer_feature_types(X)
        return X.drop(columns=self._cols_to_drop)

pipeline_instance = MulticlassClassificationPipeline(['Imputer', MyDropNullColumns,
↪Component',
                                                    'DateTime Featurization',
                                                    'Text Featurization Component',
↪Forest Classifier'],
                                                    'One Hot Encoder', 'Random',
                                                    custom_name="Pipeline with",
↪Custom Component",
                                                    random_seed=20)

code = generate_pipeline_code(pipeline_instance)
print(code)

# This string can then be pasted into a separate window and run, although since the_
↪pipeline has custom component `MyDropNullColumns`,
# the code for that component must also be included
from evalml.demos import load_fraud
X, y = load_fraud(1000)
exec(code)
pipeline.fit(X, y)

```



```

from evalml.pipelines.multiclass_classification_pipeline import
↳MulticlassClassificationPipeline
pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↳', 'y'], 'My Drop Null Columns Transformer': [MyDropNullColumns, 'Imputer.x', 'y'],
↳'DateTime Featurization Component': ['DateTime Featurization Component', 'My Drop
↳Null Columns Transformer.x', 'y'], 'Text Featurization Component': ['Text
↳Featurization Component', 'DateTime Featurization Component.x', 'y'], 'One Hot
↳Encoder': ['One Hot Encoder', 'Text Featurization Component.x', 'y'], 'Random
↳Forest Classifier': ['Random Forest Classifier', 'One Hot Encoder.x', 'y']},
↳parameters={'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_
↳impute_strategy': 'mean', 'categorical_fill_value': None, 'numeric_fill_value':
↳None}, 'My Drop Null Columns Transformer':{'pct_null_threshold': 1.0}, 'DateTime
↳Featurization Component':{'features_to_extract': ['year', 'month', 'day_of_week',
↳'hour'], 'encode_as_categories': False, 'date_index': None}, 'One Hot Encoder':{'
↳top_n': 10, 'features_to_encode': None, 'categories': None, 'drop': 'if_binary',
↳'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'Random Forest Classifier':{'
↳'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}, custom_name='Pipeline with
↳Custom Component', random_seed=20)
        Number of Features
Boolean                1
Categorical            6
Numeric                5

Number of training examples: 1000
Targets
False    85.90%
True     14.10%
Name: fraud, dtype: object

```

```

[18]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer', 'X
↳', 'y'], 'My Drop Null Columns Transformer': [MyDropNullColumns, 'Imputer.x', 'y'],
↳'DateTime Featurization Component': ['DateTime Featurization Component', 'My Drop
↳Null Columns Transformer.x', 'y'], 'Text Featurization Component': ['Text
↳Featurization Component', 'DateTime Featurization Component.x', 'y'], 'One Hot
↳Encoder': ['One Hot Encoder', 'Text Featurization Component.x', 'y'], 'Random
↳Forest Classifier': ['Random Forest Classifier', 'One Hot Encoder.x', 'y']},
↳parameters={'Imputer':{'categorical_impute_strategy': 'most_frequent', 'numeric_
↳impute_strategy': 'mean', 'categorical_fill_value': None, 'numeric_fill_value':
↳None}, 'My Drop Null Columns Transformer':{'pct_null_threshold': 1.0}, 'DateTime
↳Featurization Component':{'features_to_extract': ['year', 'month', 'day_of_week',
↳'hour'], 'encode_as_categories': False, 'date_index': None}, 'One Hot Encoder':{'
↳top_n': 10, 'features_to_encode': None, 'categories': None, 'drop': 'if_binary',
↳'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'Random Forest Classifier':{'
↳'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}, custom_name='Pipeline with
↳Custom Component', random_seed=20)

```

4.3 Component Graphs

EvalML component graphs represent and describe the flow of data in a collection of related components. A component graph is comprised of nodes representing components, and edges between pairs of nodes representing where the inputs and outputs of each component should go. It is the backbone of the features offered by the EvalML *pipeline*, but is also a powerful data structure on its own. EvalML currently supports component graphs as linear and *directed acyclic graphs* (DAG).

4.3.1 Defining a Component Graph

Component graphs can be defined by specifying the dictionary of components and edges that describe the graph.

In this dictionary, each key is a reference name for a component. Each corresponding value is a list, where the first element is the component itself, and the remaining elements are the input edges that should be connected to that component. The component as listed in the value can either be the component object itself or its string name.

This structure is very similar to that of [Dask computation graphs](#).

For example, in the code example below, we have a simple component graph made up of two components: an Imputer and a Random Forest Classifier. The names used to reference these two components are given by the keys, “My Imputer” and “RF Classifier” respectively. Each value in the dictionary is a list where the first element is the component corresponding to the component name, and the remaining elements are the inputs, e.g. “My Imputer” represents an Imputer component which has inputs “X” (the original features matrix) and “y” (the original target).

Feature edges are specified as “X” or “{component_name}.x”. For example, {“My Component”: [MyComponent, “Imputer.x”, ...]} indicates that we should use the feature output of the Imputer as part of the feature input for MyComponent. Similarly, target edges are specified as “y” or “{component_name}.y”. {“My Component”: [MyComponent, “Target Imputer.y”, ...]} indicates that we should use the target output of the Target Imputer as a target input for MyComponent.

Each component can have a number of feature inputs, but can only have one target input. All input edges must be explicitly defined.

Using a real example, we define a simple component graph consisting of three nodes: an Imputer (“My Imputer”), an One-Hot Encoder (“OHE”), and a Random Forest Classifier (“RF Classifier”).

- “My Imputer” takes the original X as a features input, and the original y as the target input
- “OHE” also takes the original X as a features input, and the original y as the target input
- “RF Classifier” takes the concatted feature outputs from “My Imputer” and “OHE” as a features input, and the original y as the target input.

```
[1]: from evalml.pipelines import ComponentGraph

component_dict = {
    'My Imputer': ['Imputer', 'X', 'y'],
    'OHE': ['One Hot Encoder', 'X', 'y'],
    'RF Classifier': ['Random Forest Classifier', 'My Imputer.x', 'OHE.x', 'y'] #_
    ↳takes in multiple feature inputs
}
cg_simple = ComponentGraph(component_dict)
```

All component graphs must end with one final or terminus node. This can either be a transformer or an estimator. Below, the component graph is invalid because has two terminus nodes: the “RF Classifier” and the “EN Classifier”.

```
[2]: # Can't instantiate a component graph with more than one terminus node (here: RF_
    ↳Classifier, EN Classifier)
component_dict = {
    'My Imputer': ['Imputer', 'X', 'y'],
    'RF Classifier': ['Random Forest Classifier', 'My Imputer.x', 'y'],
    'EN Classifier': ['Elastic Net Classifier', 'My Imputer.x', 'y']
}
```

Once we have defined a component graph, we can instantiate the graph with specific parameter values for each component using `.instantiate(parameters)`. All components in a component graph must be instantiated before fitting, transforming, or predicting.

Below, we instantiate our graph and set the value of our Imputer's `numeric_impute_strategy` to "most_frequent".

```
[3]: cg_simple.instantiate({'My Imputer': {'numeric_impute_strategy': 'most_frequent'}})
[3]: {'My Imputer': ['Imputer', 'X', 'y'], 'OHE': ['One Hot Encoder', 'X', 'y'], 'RF_
↳Classifier': ['Random Forest Classifier', 'My Imputer.x', 'OHE.x', 'y']}
```

4.3.2 Components in the Component Graph

You can use `.get_component(name)` and provide the unique component name to access any component in the component graph. Below, we can grab our Imputer component and confirm that `numeric_impute_strategy` has indeed been set to "most_frequent".

```
[4]: cg_simple.get_component('My Imputer')
[4]: Imputer(categorical_impute_strategy='most_frequent', numeric_impute_strategy='most_
↳frequent', categorical_fill_value=None, numeric_fill_value=None)
```

You can also `.get_inputs(name)` and provide the unique component name to retrieve all inputs for that component.

Below, we can grab our 'RF Classifier' component and confirm that we use "My Imputer.x" as our features input and "y" as target input.

```
[5]: cg_simple.get_inputs('RF Classifier')
[5]: ['My Imputer.x', 'OHE.x', 'y']
```

4.3.3 Component Graph Computation Order

Upon initialization, each component graph will generate a topological order. We can access this generated order by calling the `.compute_order` attribute. This attribute is used to determine the order that components should be evaluated during calls to fit and transform.

```
[6]: cg_simple.compute_order
[6]: ['OHE', 'My Imputer', 'RF Classifier']
```

4.3.4 Visualizing Component Graphs

We can get more information about an instantiated component graph by calling `.describe()`. This method will pretty-print each of the components in the graph and its parameters.

```
[7]: # Using a more involved component graph with more complex edges
component_dict = {
    "Imputer": ["Imputer", "X", "y"],
    "Target Imputer": ["Target Imputer", "X", "y"],
    "OneHot_RandomForest": ["One Hot Encoder", "Imputer.x", "Target Imputer.y"],
    "OneHot_ElasticNet": ["One Hot Encoder", "Imputer.x", "y"],
    "Random Forest": ["Random Forest Classifier", "OneHot_RandomForest.x", "y"],
    "Elastic Net": ["Elastic Net Classifier", "OneHot_ElasticNet.x", "Target_
↳Imputer.y"],
    "Logistic Regression": [
```

(continues on next page)

(continued from previous page)

```

        "Logistic Regression Classifier",
        "Random Forest.x",
        "Elastic Net.x",
        "y",
    ],
}
cg_with_estimators = ComponentGraph(component_dict)
cg_with_estimators.instantiate({})
cg_with_estimators.describe()

1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. Target Imputer
    * impute_strategy : most_frequent
    * fill_value : None
3. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
4. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
5. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6
    * n_jobs : -1
6. Elastic Net Classifier
    * penalty : elasticnet
    * C : 1.0
    * l1_ratio : 0.15
    * n_jobs : -1
    * multi_class : auto
    * solver : saga
7. Logistic Regression Classifier
    * penalty : l2
    * C : 1.0
    * n_jobs : -1
    * multi_class : auto
    * solver : lbfgs

```

We can also visualize a component graph by calling `.graph()`.

```
[8]: cg_with_estimators.graph()
```

```
[8]:
```

4.3.5 Component graph methods

Similar to the pipeline structure, we can call `fit`, `transform` or `predict`.

We can also call `fit_features` which will fit all but the final component and `compute_final_component_features` which will transform all but the final component. These two methods may be useful in cases where you want to understand what transformed features are being passed into the last component.

```
[9]: from evalml.demos import load_breast_cancer

X, y = load_breast_cancer()
component_dict = {
    'My Imputer': ['Imputer', 'X', 'y'],
    'OHE': ['One Hot Encoder', 'My Imputer.x', 'y']
}
cg_with_final_transformer = ComponentGraph(component_dict)
cg_with_final_transformer.instantiate({})
cg_with_final_transformer.fit(X, y)

# We can call `transform` for ComponentGraphs with a final transformer
cg_with_final_transformer.transform(X, y)
```

```
      Number of Features
Numeric                30
```

```
Number of training examples: 569
Targets
benign      62.74%
malignant   37.26%
Name: target, dtype: object
```

```
[9]:
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	\
0	17.99	10.38	122.80	1001.0	0.11840	
1	20.57	17.77	132.90	1326.0	0.08474	
2	19.69	21.25	130.00	1203.0	0.10960	
3	11.42	20.38	77.58	386.1	0.14250	
4	20.29	14.34	135.10	1297.0	0.10030	
..	
564	21.56	22.39	142.00	1479.0	0.11100	
565	20.13	28.25	131.20	1261.0	0.09780	
566	16.60	28.08	108.30	858.1	0.08455	
567	20.60	29.33	140.10	1265.0	0.11780	
568	7.76	24.54	47.92	181.0	0.05263	

	mean compactness	mean concavity	mean concave points	mean symmetry	\
0	0.27760	0.30010	0.14710	0.2419	
1	0.07864	0.08690	0.07017	0.1812	
2	0.15990	0.19740	0.12790	0.2069	
3	0.28390	0.24140	0.10520	0.2597	
4	0.13280	0.19800	0.10430	0.1809	
..	
564	0.11590	0.24390	0.13890	0.1726	
565	0.10340	0.14400	0.09791	0.1752	
566	0.10230	0.09251	0.05302	0.1590	
567	0.27700	0.35140	0.15200	0.2397	
568	0.04362	0.00000	0.00000	0.1587	


```
mean fractal dimension ... worst radius worst texture \
```

(continues on next page)

(continued from previous page)

```

0          0.07871 ...      25.380      17.33
1          0.05667 ...      24.990      23.41
2          0.05999 ...      23.570      25.53
3          0.09744 ...      14.910      26.50
4          0.05883 ...      22.540      16.67
..          ... ..
564        0.05623 ...      25.450      26.40
565        0.05533 ...      23.690      38.25
566        0.05648 ...      18.980      34.12
567        0.07016 ...      25.740      39.42
568        0.05884 ...       9.456      30.37

      worst perimeter  worst area  worst smoothness  worst compactness  \
0          184.60      2019.0          0.16220          0.66560
1          158.80      1956.0          0.12380          0.18660
2          152.50      1709.0          0.14440          0.42450
3           98.87       567.7          0.20980          0.86630
4          152.20      1575.0          0.13740          0.20500
..          ... ..
564        166.10      2027.0          0.14100          0.21130
565        155.00      1731.0          0.11660          0.19220
566        126.70      1124.0          0.11390          0.30940
567        184.60      1821.0          0.16500          0.86810
568         59.16       268.6          0.08996          0.06444

      worst concavity  worst concave points  worst symmetry  \
0          0.7119          0.2654          0.4601
1          0.2416          0.1860          0.2750
2          0.4504          0.2430          0.3613
3          0.6869          0.2575          0.6638
4          0.4000          0.1625          0.2364
..          ... ..
564        0.4107          0.2216          0.2060
565        0.3215          0.1628          0.2572
566        0.3403          0.1418          0.2218
567        0.9387          0.2650          0.4087
568        0.0000          0.0000          0.2871

      worst fractal dimension
0          0.11890
1          0.08902
2          0.08758
3          0.17300
4          0.07678
..          ... ..
564        0.07115
565        0.06637
566        0.07820
567        0.12400
568        0.07039

[569 rows x 30 columns]

```

```

[10]: cg_with_estimators.fit(X, y)

# We can call `predict` for ComponentGraphs with a final transformer
cg_with_estimators.predict(X)

```

```
[10]: 0      malignant
      1      malignant
      2      malignant
      3      malignant
      4      malignant
      ...
      564    malignant
      565    malignant
      566    malignant
      567    malignant
      568      benign
      Length: 569, dtype: category
      Categories (2, object): ['benign', 'malignant']
```

4.4 Components

Components are the lowest level of building blocks in EvalML. Each component represents a fundamental operation to be applied to data.

All components accept parameters as keyword arguments to their `__init__` methods. These parameters can be used to configure behavior.

Each component class definition must include a human-readable `name` for the component. Additionally, each component class may expose parameters for AutoML search by defining a `hyperparameter_ranges` attribute containing the parameters in question.

EvalML splits components into two categories: **transformers** and **estimators**.

4.4.1 Transformers

Transformers subclass the `Transformer` class, and define a `fit` method to learn information from training data and a `transform` method to apply a learned transformation to new data.

For example, an *imputer* is configured with the desired impute strategy to follow, for instance the mean value. The imputers `fit` method would learn the mean from the training data, and the `transform` method would fill the learned mean value in for any missing values in new data.

All transformers can execute `fit` and `transform` separately or in one step by calling `fit_transform`. Defining a custom `fit_transform` method can facilitate useful performance optimizations in some cases.

```
[1]: import numpy as np
      import pandas as pd
      from evalml.pipelines.components import SimpleImputer

      X = pd.DataFrame([[1, 2, 3], [1, np.nan, 3]])
      display(X)

      0      1      2
0      1      2.0    3
1      1      NaN    3
```

```
[2]: import woodwork as ww
      imp = SimpleImputer(impute_strategy="mean")

      X.ww.init()
```

(continues on next page)

(continued from previous page)

```
X = imp.fit_transform(X)
display(X)
```

```
   0    1    2
0  1  2.0  3
1  1  2.0  3
```

Below is a list of all transformers included with EvalML:

```
[3]: from evalml.pipelines.components.utils import all_components, Estimator, Transformer
     for component in all_components():
         if issubclass(component, Transformer):
             print(f"Transformer: {component.name}")
```

```
Transformer: Drop Rows Transformer
Transformer: URL Featurizer
Transformer: Email Featurizer
Transformer: DFS Transformer
Transformer: Delayed Feature Transformer
Transformer: Text Featurization Component
Transformer: LSA Transformer
Transformer: Drop Null Columns Transformer
Transformer: DateTime Featurization Component
Transformer: PCA Transformer
Transformer: Linear Discriminant Analysis Transformer
Transformer: Select Columns By Type Transformer
Transformer: Select Columns Transformer
Transformer: Drop Columns Transformer
Transformer: Oversampler
Transformer: Undersampler
Transformer: Standard Scaler
Transformer: Target Imputer
Transformer: Imputer
Transformer: Per Column Imputer
Transformer: Simple Imputer
Transformer: RF Regressor Select From Model
Transformer: RF Classifier Select From Model
Transformer: Target Encoder
Transformer: One Hot Encoder
Transformer: Log Transformer
Transformer: Polynomial Detrender
```

4.4.2 Estimators

Each estimator wraps an ML algorithm. Estimators subclass the `Estimator` class, and define a `fit` method to learn information from training data and a `predict` method for generating predictions from new data. Classification estimators should also define a `predict_proba` method for generating predicted probabilities.

Estimator classes each define a `model_family` attribute indicating what type of model is used.

Here's an example of using the *LogisticRegressionClassifier* estimator to fit and predict on a simple dataset:

```
[4]: from evalml.pipelines.components import LogisticRegressionClassifier

     clf = LogisticRegressionClassifier()
```

(continues on next page)

(continued from previous page)

```
X = X
y = [1, 0]

clf.fit(X, y)
clf.predict(X)
```

```
[4]: 0      0
      1      0
      dtype: int64
```

Below is a list of all estimators included with EvalML:

```
[5]: from evalml.pipelines.components.utils import all_components, Estimator, Transformer
      for component in all_components():
          if issubclass(component, Estimator):
              print(f"Estimator: {component.name}")
```

```
Estimator: Sklearn Stacked Ensemble Regressor
Estimator: Sklearn Stacked Ensemble Classifier
Estimator: ARIMA Regressor
Estimator: SVM Regressor
Estimator: Time Series Baseline Estimator
Estimator: Decision Tree Regressor
Estimator: Baseline Regressor
Estimator: Extra Trees Regressor
Estimator: XGBoost Regressor
Estimator: CatBoost Regressor
Estimator: Random Forest Regressor
Estimator: LightGBM Regressor
Estimator: Linear Regressor
Estimator: Elastic Net Regressor
Estimator: SVM Classifier
Estimator: KNN Classifier
Estimator: Decision Tree Classifier
Estimator: LightGBM Classifier
Estimator: Baseline Classifier
Estimator: Extra Trees Classifier
Estimator: Elastic Net Classifier
Estimator: CatBoost Classifier
Estimator: XGBoost Classifier
Estimator: Random Forest Classifier
Estimator: Logistic Regression Classifier
```

4.4.3 Defining Custom Components

EvalML allows you to easily create your own custom components by following the steps below.

Custom Transformers

Your transformer must inherit from the correct subclass. In this case *Transformer* for components that transform data. Next we will use EvalML's *DropNullColumns* as an example.

```
[6]: from evalml.pipelines.components import Transformer
from evalml.utils import (
    infer_feature_types,
)

class DropNullColumns(Transformer):
    """Transformer to drop features whose percentage of NaN values exceeds a
    specified threshold"""
    name = "Drop Null Columns Transformer"
    hyperparameter_ranges = {}

    def __init__(self, pct_null_threshold=1.0, random_seed=0, **kwargs):
        """Initializes an transformer to drop features whose percentage of NaN values
        exceeds a specified threshold.

        Arguments:
            pct_null_threshold(float): The percentage of NaN values in an input
            feature to drop.
            Must be a value between [0, 1] inclusive. If equal to 0.0, will drop
            columns with any null values.
            If equal to 1.0, will drop columns with all null values. Defaults to
            0.95.
        """
        if pct_null_threshold < 0 or pct_null_threshold > 1:
            raise ValueError("pct_null_threshold must be a float between 0 and 1,
            inclusive.")
        parameters = {"pct_null_threshold": pct_null_threshold}
        parameters.update(kwargs)

        self._cols_to_drop = None
        super().__init__(parameters=parameters,
                         component_obj=None,
                         random_seed=random_seed)

    def fit(self, X, y=None):
        """Fits DropNullColumns component to data

        Arguments:
            X (pd.DataFrame): The input training data of shape [n_samples, n_features]
            y (pd.Series, optional): The target training data of length [n_samples]

        Returns:
            self
        """
        pct_null_threshold = self.parameters["pct_null_threshold"]
        X_t = infer_feature_types(X)
        percent_null = X_t.isnull().mean()
        if pct_null_threshold == 0.0:
```

(continues on next page)

(continued from previous page)

```

        null_cols = percent_null[percent_null > 0]
    else:
        null_cols = percent_null[percent_null >= pct_null_threshold]
    self._cols_to_drop = list(null_cols.index)
    return self

    def transform(self, X, y=None):
        """Transforms data X by dropping columns that exceed the threshold of null_
        ↪ values.

        Arguments:
            X (pd.DataFrame): Data to transform
            y (pd.Series, optional): Ignored.

        Returns:
            pd.DataFrame: Transformed X
        """
        X_t = infer_feature_types(X)
        return X_t.drop(self._cols_to_drop)

```

Required fields

- `name`: A human-readable name.
- `modifies_features`: A boolean that specifies whether this component modifies (subsets or transforms) the features variable during `transform`.
- `modifies_target`: A boolean that specifies whether this component modifies (subsets or transforms) the target variable during `transform`.

Required methods

Likewise, there are select methods you need to override as `Transformer` is an abstract base class:

- `__init__()`: The `__init__()` method of your transformer will need to call `super().__init__()` and pass three parameters in: a `parameters` dictionary holding the parameters to the component, the `component_obj`, and the `random_seed` value. You can see that `component_obj` is set to `None` above and we will discuss `component_obj` in depth later on.
- `fit()`: The `fit()` method is responsible for fitting your component on training data. It should return the component object.
- `transform()`: After fitting a component, the `transform()` method will take in new data and transform accordingly. It should return a pandas dataframe with `woodwork` initialized. Note: a component must call `fit()` before `transform()`.

You can also call or override `fit_transform()` that combines `fit()` and `transform()` into one method.

Custom Estimators

Your estimator must inherit from the correct subclass. In this case *Estimator* for components that predict new target values. Next we will use EvalML's *BaselineRegressor* as an example.

```
[7]: import numpy as np
import pandas as pd

from evalml.model_family import ModelFamily
from evalml.pipelines.components.estimators import Estimator
from evalml.problem_types import ProblemTypes

class BaselineRegressor(Estimator):
    """Regressor that predicts using the specified strategy.

    This is useful as a simple baseline regressor to compare with other regressors.
    """
    name = "Baseline Regressor"
    hyperparameter_ranges = {}
    model_family = ModelFamily.BASELINE
    supported_problem_types = [ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_
    ↪ REGRESSION]

    def __init__(self, strategy="mean", random_seed=0, **kwargs):
        """Baseline regressor that uses a simple strategy to make predictions.

        Arguments:
            strategy (str): Method used to predict. Valid options are "mean", "median
            ↪ ". Defaults to "mean".
            random_seed (int): Seed for the random number generator. Defaults to 0.

        """
        if strategy not in ["mean", "median"]:
            raise ValueError("'strategy' parameter must equal either 'mean' or 'median
            ↪ ")
        parameters = {"strategy": strategy}
        parameters.update(kwargs)

        self._prediction_value = None
        self._num_features = None
        super().__init__(parameters=parameters,
                         component_obj=None,
                         random_seed=random_seed)

    def fit(self, X, y=None):
        if y is None:
            raise ValueError("Cannot fit Baseline regressor if y is None")
        X = infer_feature_types(X)
        y = infer_feature_types(y)

        if self.parameters["strategy"] == "mean":
            self._prediction_value = y.mean()
        elif self.parameters["strategy"] == "median":
            self._prediction_value = y.median()
        self._num_features = X.shape[1]
        return self
```

(continues on next page)

(continued from previous page)

```

def predict(self, X):
    X = infer_feature_types(X)
    predictions = pd.Series([self._prediction_value] * len(X))
    return infer_feature_types(predictions)

@property
def feature_importance(self):
    """Returns importance associated with each feature. Since baseline regressors
    ↪do not use input features to calculate predictions, returns an array of zeroes.

    Returns:
        np.ndarray (float): An array of zeroes

    """
    return np.zeros(self._num_features)

```

Required fields

- name: A human-readable name.
- model_family - EvalML *model_family* that this component belongs to
- supported_problem_types - list of EvalML *problem_types* that this component supports
- modifies_features: A boolean that specifies whether the return value from predict or predict_proba should be used as features.
- modifies_target: A boolean that specifies whether the return value from predict or predict_proba should be used as the target variable.

Model families and problem types include:

```

[8]: from evalml.model_family import ModelFamily
    from evalml.problem_types import ProblemTypes

print("Model Families:\n", [m.value for m in ModelFamily])
print("Problem Types:\n", [p.value for p in ProblemTypes])

Model Families:
['k_neighbors', 'random_forest', 'svm', 'xgboost', 'lightgbm', 'linear_model',
↪'catboost', 'extra_trees', 'ensemble', 'decision_tree', 'arima', 'baseline',
↪'prophet', 'none']
Problem Types:
['binary', 'multiclass', 'regression', 'time series regression', 'time series binary',
↪, 'time series multiclass']

```

Required methods

- `__init__()` - the `__init__()` method of your estimator will need to call `super().__init__()` and pass three parameters in: a `parameters` dictionary holding the parameters to the component, the `component_obj`, and the `random_seed` value.
- `fit()` - the `fit()` method is responsible for fitting your component on training data.
- `predict()` - after fitting a component, the `predict()` method will take in new data and predict new target values. Note: a component must call `fit()` before `predict()`.
- `feature_importance` - `feature_importance` is a [Python property](#) that returns a list of importances associated with each feature.

If your estimator handles classification problems it also requires an additional method:

- `predict_proba()` - this method predicts probability estimates for classification labels

Components Wrapping Third-Party Objects

The `component_obj` parameter is used for wrapping third-party objects and using them in component implementation. If you're using a `component_obj` you will need to define `__init__()` and pass in the relevant object that has also implemented the required methods mentioned above. However, if the `component_obj` does not follow EvalML component conventions, you may need to override methods as needed. Below is an example of EvalML's *LinearRegressor*.

```
[9]: from sklearn.linear_model import LinearRegression as SKLinearRegression

from evalml.model_family import ModelFamily
from evalml.pipelines.components.estimators import Estimator
from evalml.problem_types import ProblemTypes

class LinearRegressor(Estimator):
    """Linear Regressor."""
    name = "Linear Regressor"
    model_family = ModelFamily.LINEAR_MODEL
    supported_problem_types = [ProblemTypes.REGRESSION]

    def __init__(self, fit_intercept=True, normalize=False, n_jobs=-1, random_seed=0,
↳ **kwargs):
        parameters = {
            'fit_intercept': fit_intercept,
            'normalize': normalize,
            'n_jobs': n_jobs
        }
        parameters.update(kwargs)
        linear_regressor = SKLinearRegression(**parameters)
        super().__init__(parameters=parameters,
                        component_obj=linear_regressor,
                        random_seed=random_seed)

    @property
    def feature_importance(self):
        return self._component_obj.coef_
```

Hyperparameter Ranges for AutoML

`hyperparameter_ranges` is a dictionary mapping the parameter name (str) to an allowed range (SkOpt Space) for that parameter. Both lists and `skopt.space.Categorical` values are accepted for categorical spaces.

AutoML will perform a search over the allowed ranges for each parameter to select models which produce optimal performance within those ranges. AutoML gets the allowed ranges for each component from the component's `hyperparameter_ranges` class attribute. Any component parameter you add an entry for in `hyperparameter_ranges` will be included in the AutoML search. If parameters are omitted, AutoML will use the default value in all pipelines.

4.4.4 Generate Component Code

Once you have a component defined in EvalML, you can generate string Python code to recreate this component, which can then be saved and run elsewhere with EvalML. `generate_component_code` requires a component instance as the input. This method works for custom components as well, although it won't return the code required to define the custom component.

```
[10]: from evalml.pipelines.components import LogisticRegressionClassifier
      from evalml.pipelines.components.utils import generate_component_code

      lr = LogisticRegressionClassifier(C=5)
      code = generate_component_code(lr)
      print(code)

      from evalml.pipelines.components.estimators.classifiers.logistic_regression_
      ↪ classifier import LogisticRegressionClassifier

      logisticRegressionClassifier = LogisticRegressionClassifier(**{'penalty': 'l2', 'C':
      ↪ 5, 'n_jobs': -1, 'multi_class': 'auto', 'solver': 'lbfgs'})

[11]: # this string can then be copy and pasted into a separate window and executed as
      ↪ python code
      exec(code)

[12]: # We can also do this for custom components
      from evalml.pipelines.components.utils import generate_component_code

      myDropNull = DropNullColumns()
      print(generate_component_code(myDropNull))

      dropNullColumnsTransformer = DropNullColumns(**{'pct_null_threshold': 1.0})
```

Expectations for Custom Classification Components

EvalML expects the following from custom classification component implementations:

- Classification targets will range from 0 to n-1 and are integers.
- For classification estimators, the order of `predict_proba`'s columns must match the order of the target, and the column names must be integers ranging from 0 to n-1

4.5 Objectives

4.5.1 Overview

One of the key choices to make when training an ML model is what metric to choose by which to measure the efficacy of the model at learning the signal. Such metrics are useful for comparing how well the trained models generalize to new similar data.

This choice of metric is a key component of AutoML because it defines the cost function the AutoML search will seek to optimize. In EvalML, these metrics are called **objectives**. AutoML will seek to minimize (or maximize) the objective score as it explores more pipelines and parameters and will use the feedback from scoring pipelines to tune the available hyperparameters and continue the search. Therefore, it is critical to have an objective function that represents how the model will be applied in the intended domain of use.

EvalML supports a variety of objectives from traditional supervised ML including [mean squared error](#) for regression problems and [cross entropy](#) or [area under the ROC curve](#) for classification problems. EvalML also allows the user to define a custom objective using their domain expertise, so that AutoML can search for models which provide the most value for the user's problem.

4.5.2 Core Objectives

Use the `get_core_objectives` method to get a list of which objectives are included with EvalML for each problem type:

```
[1]: from evalml.objectives import get_core_objectives
from evalml.problem_types import ProblemTypes

for objective in get_core_objectives(ProblemTypes.BINARY):
    print(objective.name)
```

```
MCC Binary
Log Loss Binary
Gini
AUC
Precision
F1
Balanced Accuracy Binary
Accuracy Binary
```

EvalML defines a base objective class for each problem type: `RegressionObjective`, `BinaryClassificationObjective` and `MulticlassClassificationObjective`. All EvalML objectives are a subclass of one of these.

Binary Classification Objectives and Thresholds

All binary classification objectives have a `threshold` property. Some binary classification objectives like log loss and AUC are unaffected by the choice of binary classification threshold, because they score based on predicted probabilities or examine a range of threshold values. These metrics are defined with `score_needs_proba` set to `False`. For all other binary classification objectives, we can compute the optimal binary classification threshold from the predicted probabilities and the target.

```
[2]: from evalml.pipelines import BinaryClassificationPipeline
from evalml.demos import load_fraud
from evalml.objectives import F1
```

(continues on next page)

(continued from previous page)

```

X, y = load_fraud(n_rows=100)
X.ww.init(logical_types={"provider": "Categorical", "region": "Categorical",
                        "currency": "Categorical", "expiration_date": "Categorical"})
objective = F1()
pipeline = BinaryClassificationPipeline(component_graph=['Simple Imputer', 'DateTime_
↳Featurization Component', 'One Hot Encoder', 'Random Forest Classifier'])
pipeline.fit(X, y)
print(pipeline.threshold)
print(pipeline.score(X, y, objectives=[objective]))

y_pred_proba = pipeline.predict_proba(X)[True]
pipeline.threshold = objective.optimize_threshold(y_pred_proba, y)
print(pipeline.threshold)
print(pipeline.score(X, y, objectives=[objective]))

```

```

                Number of Features
Boolean                      1
Categorical                   6
Numeric                       5

Number of training examples: 100
Targets
False      91.00%
True       9.00%
Name: fraud, dtype: object
None
OrderedDict([('F1', 1.0)])
0.5202757772593112
OrderedDict([('F1', 1.0)])

```

4.5.3 Custom Objectives

Often times, the objective function is very specific to the use-case or business problem. To get the right objective to optimize requires thinking through the decisions or actions that will be taken using the model and assigning a cost/benefit to doing that correctly or incorrectly based on known outcomes in the training data.

Once you have determined the objective for your business, you can provide that to EvalML to optimize by defining a custom objective function.

Defining a Custom Objective Function

To create a custom objective class, we must define several elements:

- **name:** The printable name of this objective.
- **objective_function:** This function takes the predictions, true labels, and an optional reference to the inputs, and returns a score of how well the model performed.
- **greater_is_better:** True if a higher `objective_function` value represents a better solution, and otherwise False.
- **score_needs_proba:** Only for classification objectives. True if the objective is intended to function with predicted probabilities as opposed to predicted values (example: cross entropy for classifiers).
- **decision_function:** Only for binary classification objectives. This function takes predicted probabilities that were output from the model and a binary classification threshold, and returns predicted values.

- `perfect_score`: The score achieved by a perfect model on this objective.
- `expected_range`: The expected range of values we want this objective to output, which doesn't necessarily have to be equal to the possible range of values. For example, our expected R2 range is from `[-1, 1]`, although the actual range is `(-inf, 1]`.

Example: Fraud Detection

To give a concrete example, let's look at how the *fraud detection* objective function is built.

```
[3]: from evalml.objectives.binary_classification_objective import _
      BinaryClassificationObjective
      import pandas as pd

class FraudCost(BinaryClassificationObjective):
    """Score the percentage of money lost of the total transaction amount process due
    to fraud"""
    name = "Fraud Cost"
    greater_is_better = False
    score_needs_proba = False
    perfect_score = 0.0

    def __init__(self, retry_percentage=.5, interchange_fee=.02,
                  fraud_payout_percentage=1.0, amount_col='amount'):
        """Create instance of FraudCost

        Arguments:
            retry_percentage (float): What percentage of customers that will retry a
            transaction if it
            is declined. Between 0 and 1. Defaults to .5

            interchange_fee (float): How much of each successful transaction you can
            collect.
            Between 0 and 1. Defaults to .02

            fraud_payout_percentage (float): Percentage of fraud you will not be able
            to collect.
            Between 0 and 1. Defaults to 1.0

            amount_col (str): Name of column in data that contains the amount.
            Defaults to "amount"
        """
        self.retry_percentage = retry_percentage
        self.interchange_fee = interchange_fee
        self.fraud_payout_percentage = fraud_payout_percentage
        self.amount_col = amount_col

    def decision_function(self, ypred_proba, threshold=0.0, X=None):
        """Determine if a transaction is fraud given predicted probabilities,
        threshold, and dataframe with transaction amount

        Arguments:
            ypred_proba (pd.Series): Predicted probabilities
            X (pd.DataFrame): Dataframe containing transaction amount
            threshold (float): Dollar threshold to determine if transaction is
            fraud
```

(continues on next page)

(continued from previous page)

```

        Returns:
            pd.Series: Series of predicted fraud labels using X and threshold
    """
    if not isinstance(X, pd.DataFrame):
        X = pd.DataFrame(X)

    if not isinstance(ypred_proba, pd.Series):
        ypred_proba = pd.Series(ypred_proba)

    transformed_probs = (ypred_proba.values * X[self.amount_col])
    return transformed_probs > threshold

def objective_function(self, y_true, y_predicted, X):
    """Calculate amount lost to fraud per transaction given predictions, true_
    values, and dataframe with transaction amount

    Arguments:
        y_predicted (pd.Series): predicted fraud labels
        y_true (pd.Series): true fraud labels
        X (pd.DataFrame): dataframe with transaction amounts

    Returns:
        float: amount lost to fraud per transaction
    """
    if not isinstance(X, pd.DataFrame):
        X = pd.DataFrame(X)

    if not isinstance(y_predicted, pd.Series):
        y_predicted = pd.Series(y_predicted)

    if not isinstance(y_true, pd.Series):
        y_true = pd.Series(y_true)

    # extract transaction using the amount columns in users data
    try:
        transaction_amount = X[self.amount_col]
    except KeyError:
        raise ValueError("`{}` is not a valid column in X.".format(self.amount_
        col))

    # amount paid if transaction is fraud
    fraud_cost = transaction_amount * self.fraud_payout_percentage

    # money made from interchange fees on transaction
    interchange_cost = transaction_amount * (1 - self.retry_percentage) * self.
    interchange_fee

    # calculate cost of missing fraudulent transactions
    false_negatives = (y_true & ~y_predicted) * fraud_cost

    # calculate money lost from fees
    false_positives = (~y_true & y_predicted) * interchange_cost

    loss = false_negatives.sum() + false_positives.sum()

    loss_per_total_processed = loss / transaction_amount.sum()

```

(continues on next page)

(continued from previous page)

```
return loss_per_total_processed
```

4.6 Model Understanding

Simply examining a model’s performance metrics is not enough to select a model and promote it for use in a production setting. While developing an ML algorithm, it is important to understand how the model behaves on the data, to examine the key factors influencing its predictions and to consider where it may be deficient. Determination of what “success” may mean for an ML project depends first and foremost on the user’s domain expertise.

EvalML includes a variety of tools for understanding models, from graphing utilities to methods for explaining predictions.

** Graphing methods on Jupyter Notebook and Jupyter Lab require `ipywidgets` to be installed.

** If graphing on Jupyter Lab, `jupyterlab-plotly` required. To download this, make sure you have `npm` installed.

4.6.1 Graphing Utilities

First, let’s train a pipeline on some data.

```
[1]: import evalml
from evalml.pipelines import BinaryClassificationPipeline
X, y = evalml.demos.load_breast_cancer()

X_train, X_holdout, y_train, y_holdout = evalml.preprocessing.split_data(X, y,
    ↪problem_type='binary',
    ↪2, random_seed=0)

pipeline_binary = BinaryClassificationPipeline(['Simple Imputer', 'Random Forest_
    ↪Classifier'])
pipeline_binary.fit(X_train, y_train)
print(pipeline_binary.score(X_holdout, y_holdout, objectives=['log loss binary']))
```

```

      Number of Features
Numeric                30

Number of training examples: 569
Targets
benign          62.74%
malignant       37.26%
Name: target, dtype: object
OrderedDict([('Log Loss Binary', 0.1686746297113362)])
```

Feature Importance

We can get the importance associated with each feature of the resulting pipeline

```
[2]: pipeline_binary.feature_importance
```

```
[2]:
```

	feature	importance
0	mean concave points	0.138857
1	worst perimeter	0.137780
2	worst concave points	0.117782
3	worst radius	0.100584
4	mean concavity	0.086402
5	worst area	0.072027
6	mean perimeter	0.046500
7	worst concavity	0.043408
8	mean radius	0.037664
9	mean area	0.033683
10	radius error	0.025036
11	area error	0.019324
12	worst texture	0.014754
13	worst compactness	0.014462
14	mean texture	0.013856
15	worst smoothness	0.013710
16	worst symmetry	0.011395
17	perimeter error	0.010284
18	mean compactness	0.008162
19	mean smoothness	0.008154
20	worst fractal dimension	0.007034
21	fractal dimension error	0.005502
22	compactness error	0.004953
23	smoothness error	0.004728
24	texture error	0.004384
25	symmetry error	0.004250
26	mean fractal dimension	0.004164
27	concavity error	0.004089
28	mean symmetry	0.003997
29	concave points error	0.003076

We can also create a bar plot of the feature importances

```
[3]: pipeline_binary.graph_feature_importance()
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Permutation Importance

We can also compute and plot the permutation importance of the pipeline.

```
[4]: from evalml.model_understanding import calculate_permutation_importance
calculate_permutation_importance(pipeline_binary, X_holdout, y_holdout, 'log loss_
↪binary')
```

```
[4]:
```

	feature	importance
0	worst perimeter	0.063657
1	worst area	0.045759

(continues on next page)

(continued from previous page)

```

2         worst radius      0.041926
3     mean concave points    0.029325
4     worst concave points   0.021045
5         worst concavity    0.010105
6         worst texture       0.010044
7         mean texture        0.006178
8         mean symmetry       0.005857
9         mean area          0.004745
10        worst smoothness    0.003190
11        area error          0.003113
12        mean perimeter      0.002478
13    mean fractal dimension   0.001981
14        compactness error    0.001968
15        concavity error      0.001947
16        texture error        0.000291
17        smoothness error     -0.000206
18        mean smoothness     -0.000745
19    fractal dimension error   -0.000835
20        worst compactness    -0.002392
21        mean concavity       -0.003188
22        mean compactness     -0.005377
23            radius error     -0.006229
24        mean radius         -0.006870
25    worst fractal dimension   -0.007415
26        symmetry error       -0.008175
27        perimeter error      -0.008980
28        concave points error -0.010415
29        worst symmetry       -0.018645

```

```
[5]: from evalml.model_understanding import graph_permutation_importance
graph_permutation_importance(pipeline_binary, X_holdout, y_holdout, 'log loss binary')
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Partial Dependence Plots

We can calculate the one-way [partial dependence plots](#) for a feature.

```
[6]: from evalml.model_understanding.graphs import partial_dependence
partial_dependence(pipeline_binary, X_holdout, features='mean radius', grid_
↳ resolution=5)
```

```
[6]:
```

	feature_values	partial_dependence	class_label
0	9.69092	0.392453	malignant
1	12.40459	0.395962	malignant
2	15.11826	0.417396	malignant
3	17.83193	0.429542	malignant
4	20.54560	0.429717	malignant

```
[7]: from evalml.model_understanding.graphs import graph_partial_dependence
graph_partial_dependence(pipeline_binary, X_holdout, features='mean radius', grid_
↳ resolution=5)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

You can also compute the partial dependence for a categorical feature. We will demonstrate this on the fraud dataset.

```
[8]: X_fraud, y_fraud = evalml.demos.load_fraud(100, verbose=False)
X_fraud.ww.init(logical_types={"provider": "Categorical", 'region': "Categorical",
                               "currency": "Categorical", "expiration_date":
                               ↪ "Categorical"})

fraud_pipeline = BinaryClassificationPipeline(["DateTime Featurization Component",
↪ "One Hot Encoder", "Random Forest Classifier"])
fraud_pipeline.fit(X_fraud, y_fraud)

graph_partial_dependence(fraud_pipeline, X_fraud, features='provider')
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Two-way partial dependence plots are also possible and invoke the same API.

```
[9]: partial_dependence(pipeline_binary, X_holdout, features=('worst_perimeter', 'worst_
↪ radius'), grid_resolution=5)
```

```
[9]:
```

	10.6876	14.404924999999999	18.12225	21.839575	25.5569	\
69.140700	0.279038		0.282898	0.435179	0.435355	0.435355
94.334275	0.304335		0.308194	0.458283	0.458458	0.458458
119.527850	0.464455		0.468314	0.612137	0.616932	0.616932
144.721425	0.483437		0.487297	0.631120	0.635915	0.635915
169.915000	0.483437		0.487297	0.631120	0.635915	0.635915
	class_label					
69.140700	malignant					
94.334275	malignant					
119.527850	malignant					
144.721425	malignant					
169.915000	malignant					

```
[10]: graph_partial_dependence(pipeline_binary, X_holdout, features=('worst_perimeter',
↪ 'worst radius'), grid_resolution=5)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Confusion Matrix

For binary or multiclass classification, we can view a [confusion matrix](#) of the classifier's predictions. In the DataFrame output of `confusion_matrix()`, the column header represents the predicted labels while row header represents the actual labels.

```
[11]: from evalml.model_understanding.graphs import confusion_matrix
y_pred = pipeline_binary.predict(X_holdout)
confusion_matrix(y_holdout, y_pred)
```

```
[11]:      benign  malignant
benign    0.930556   0.069444
malignant  0.023810   0.976190
```

```
[12]: from evalml.model_understanding.graphs import graph_confusion_matrix
y_pred = pipeline_binary.predict(X_holdout)
graph_confusion_matrix(y_holdout, y_pred)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Precision-Recall Curve

For binary classification, we can view the precision-recall curve of the pipeline.

```
[13]: from evalml.model_understanding.graphs import graph_precision_recall_curve
# get the predicted probabilities associated with the "true" label
import woodwork as ww
y_encoded = y_holdout.ww.map({'benign': 0, 'malignant': 1})
y_pred_proba = pipeline_binary.predict_proba(X_holdout)["malignant"]
graph_precision_recall_curve(y_encoded, y_pred_proba)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

ROC Curve

For binary and multiclass classification, we can view the [Receiver Operating Characteristic \(ROC\)](#) curve of the pipeline.

```
[14]: from evalml.model_understanding.graphs import graph_roc_curve
# get the predicted probabilities associated with the "malignant" label
y_pred_proba = pipeline_binary.predict_proba(X_holdout)["malignant"]
graph_roc_curve(y_encoded, y_pred_proba)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

The ROC curve can also be generated for multiclass classification problems. For multiclass problems, the graph will show a one-vs-many ROC curve for each class.


```
[15]: from evalml.pipelines import MulticlassClassificationPipeline
X_multi, y_multi = evalml.demos.load_wine()

pipeline_multi = MulticlassClassificationPipeline(['Simple Imputer', 'Random Forest_
↳Classifier'])
pipeline_multi.fit(X_multi, y_multi)

y_pred_proba = pipeline_multi.predict_proba(X_multi)
graph_roc_curve(y_multi, y_pred_proba)
```

```

      Number of Features
Numeric                13

Number of training examples: 178
Targets
class_1    39.89%
class_0    33.15%
class_2    26.97%
Name: target, dtype: object
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Binary Objective Score vs. Threshold Graph

Some binary classification objectives (objectives that have `score_needs_proba` set to `False`) are sensitive to a decision threshold. For those objectives, we can obtain and graph the scores for thresholds from zero to one, calculated at evenly-spaced intervals determined by `steps`.

```
[16]: from evalml.model_understanding.graphs import binary_objective_vs_threshold
binary_objective_vs_threshold(pipeline_binary, X_holdout, y_holdout, 'f1', steps=10)
```

```
[16]:
threshold    score
0          0.0  0.538462
1          0.1  0.811881
2          0.2  0.891304
3          0.3  0.901099
4          0.4  0.931818
5          0.5  0.931818
6          0.6  0.941176
7          0.7  0.951220
8          0.8  0.936709
9          0.9  0.923077
10         1.0  0.000000
```

```
[17]: from evalml.model_understanding.graphs import graph_binary_objective_vs_threshold
graph_binary_objective_vs_threshold(pipeline_binary, X_holdout, y_holdout, 'f1',
↳steps=100)
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Predicted Vs Actual Values Graph for Regression Problems

We can also create a scatterplot comparing predicted vs actual values for regression problems. We can specify an `outlier_threshold` to color values differently if the absolute difference between the actual and predicted values are outside of a given threshold.

```
[18]: from evalml.model_understanding.graphs import graph_prediction_vs_actual
from evalml.pipelines import RegressionPipeline

X_regress, y_regress = evalml.demos.load_diabetes()
X_train, X_test, y_train, y_test = evalml.preprocessing.split_data(X_regress, y_
→regress, problem_type='regression')

pipeline_regress = RegressionPipeline(['One Hot Encoder', 'Linear Regressor'])
pipeline_regress.fit(X_train, y_train)

y_pred = pipeline_regress.predict(X_test)
graph_prediction_vs_actual(y_test, y_pred, outlier_threshold=50)
```

```

      Number of Features
Numeric                10

Number of training examples: 442
Targets
200.0    1.36%
 72.0    1.36%
 90.0    1.13%
178.0    1.13%
 71.0    1.13%
...
 73.0    0.23%
222.0    0.23%
 86.0    0.23%
 79.0    0.23%
 57.0    0.23%
Name: target, Length: 214, dtype: object
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

Now let's train a decision tree on some data.

```
[19]: pipeline_dt = BinaryClassificationPipeline(['Simple Imputer', 'Decision Tree_
→Classifier'])
pipeline_dt.fit(X_train, y_train)

[19]: pipeline = BinaryClassificationPipeline(component_graph={'Simple Imputer': ['Simple_
→Imputer', 'X', 'y'], 'Decision Tree Classifier': ['Decision Tree Classifier',
→'Simple Imputer.x', 'y']}, parameters={'Simple Imputer':{'impute_strategy': 'most_
→frequent', 'fill_value': None}, 'Decision Tree Classifier':{'criterion': 'gini',
→'max_features': 'auto', 'max_depth': 6, 'min_samples_split': 2, 'min_weight_
→fraction_leaf': 0.0}}, random_seed=0)
```

Tree Visualization

We can visualize the structure of the Decision Tree that was fit to that data, and save it if necessary.

```
[20]: from evalml.model_understanding.graphs import visualize_decision_tree

visualize_decision_tree(pipeline_dt.estimated, max_depth=2, rotate=False, filled=True,
↳ filepath=None)
```

[20]:

4.6.2 Explaining Predictions

We can explain why the model made certain predictions with the `explain_predictions` function. This will use the [Shapley Additive Explanations \(SHAP\)](#) algorithm to identify the top features that explain the predicted value.

This function can explain both classification and regression models - all you need to do is provide the pipeline, the input features, and a list of rows corresponding to the indices of the input features you want to explain. The function will return a table that you can print summarizing the top 3 most positive and negative contributing features to the predicted value.

In the example below, we explain the prediction for the third data point in the data set. We see that the worst concave points feature increased the estimated probability that the tumor is malignant by 20% while the worst radius feature decreased the probability the tumor is malignant by 5%.

```
[21]: from evalml.model_understanding.prediction_explanations import explain_predictions

table = explain_predictions(pipeline=pipeline_binary, input_features=X_holdout,
↳ y=None, indices_to_explain=[3],
                                top_k_features=6, include_shap_values=True)

print(table)
```

Random Forest Classifier w/ Simple Imputer

```
{'Simple Imputer': {'impute_strategy': 'most_frequent', 'fill_value': None}, 'Random_
↳ Forest Classifier': {'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}
```

1 of 1

	Feature Name	Feature Value	Contribution to Prediction	
↳ SHAP Value				
↳ =====				
↳ 0.02	worst concavity	0.18	-	-
↳ 0.03	mean concavity	0.04	-	-
↳ 0.03	worst area	599.50	-	-
↳ 0.05	worst radius	14.04	-	-
↳ 0.05	mean concave points	0.03	-	-
↳ 0.06	worst perimeter	92.80	-	-

The interpretation of the table is the same for regression problems - but the SHAP value now corresponds to the change in the estimated value of the dependent variable rather than a change in probability. For multiclass classification problems, a table will be output for each possible class.

Below is an example of how you would explain three predictions with *explain_predictions*.

```
[22]: from evalml.model_understanding.prediction_explanations import explain_predictions

report = explain_predictions(pipeline=pipeline_binary,
                             input_features=X_holdout, y=y_holdout, indices_to_
                             ↪explain=[0, 4, 9], include_shap_values=True,
                             output_format='text')

print(report)
```

Random Forest Classifier w/ Simple Imputer

```
{'Simple Imputer': {'impute_strategy': 'most_frequent', 'fill_value': None}, 'Random_
↪Forest Classifier': {'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}
```

1 of 3

↪SHAP Value	Feature Name	Feature Value	Contribution to Prediction
↪-0.04	worst perimeter	101.20	-
↪-0.05	worst concave points	0.06	-
↪-0.05	mean concave points	0.01	-

2 of 3

↪SHAP Value	Feature Name	Feature Value	Contribution to Prediction
↪0.05	worst radius	11.94	-
↪0.06	worst perimeter	80.78	-
↪0.06	mean concave points	0.02	-

3 of 3

↪SHAP Value	Feature Name	Feature Value	Contribution to Prediction
↪-0.05	worst concave points	0.10	-
↪-0.06	worst perimeter	99.21	-
↪-0.08	mean concave points	0.03	-

(continues on next page)

(continued from previous page)

Explaining Best and Worst Predictions

When debugging machine learning models, it is often useful to analyze the best and worst predictions the model made. The `explain_predictions_best_worst` function can help us with this.

This function will display the output of `explain_predictions` for the best 2 and worst 2 predictions. By default, the best and worst predictions are determined by the absolute error for regression problems and `cross entropy` for classification problems.

We can specify our own ranking function by passing in a function to the `metric` parameter. This function will be called on `y_true` and `y_pred`. By convention, lower scores are better.

At the top of each table, we can see the predicted probabilities, target value, error, and row index for that prediction. For a regression problem, we would see the predicted value instead of predicted probabilities.

```
[23]: from evalml.model_understanding.prediction_explanations import explain_predictions_
      ↪ best_worst

report = explain_predictions_best_worst(pipeline=pipeline_binary, input_features=X_
      ↪ holdout, y_true=y_holdout,
                                      include_shap_values=True, top_k_features=6,
      ↪ num_to_explain=2)

print(report)
```

Random Forest Classifier w/ Simple Imputer

```
{'Simple Imputer': {'impute_strategy': 'most_frequent', 'fill_value': None}, 'Random_
      ↪ Forest Classifier': {'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}
```

Best 1 of 2

```
Predicted Probabilities: [benign: 1.0, malignant: 0.0]
Predicted Value: benign
Target Value: benign
Cross Entropy: 0.0
Index ID: 502
```

	Feature Name	Feature Value	Contribution to Prediction
↪ SHAP Value			
↪	=====		
↪ -0.03	mean concavity	0.06	-
↪ -0.03	worst area	552.00	-
↪ -0.05	worst concave points	0.08	-
↪ -0.05	worst radius	13.57	-
↪ -0.05	mean concave points	0.03	-

(continues on next page)

(continued from previous page)

↪-0.06	worst perimeter	86.67	-	↪
Best 2 of 2				
Predicted Probabilities: [benign: 1.0, malignant: 0.0]				
Predicted Value: benign				
Target Value: benign				
Cross Entropy: 0.0				
Index ID: 52				
↪SHAP Value	Feature Name	Feature Value	Contribution to Prediction	↪
↪=====	=====			
↪-0.02	mean concavity	0.02	-	↪
↪-0.03	worst area	527.20	-	↪
↪-0.04	worst radius	13.10	-	↪
↪-0.04	worst concave points	0.06	-	↪
↪-0.05	mean concave points	0.01	-	↪
↪-0.06	worst perimeter	83.67	-	↪
Worst 1 of 2				
Predicted Probabilities: [benign: 0.266, malignant: 0.734]				
Predicted Value: malignant				
Target Value: benign				
Cross Entropy: 1.325				
Index ID: 363				
↪Value	Feature Name	Feature Value	Contribution to Prediction	SHAP↪
↪=====	=====			
	worst perimeter	117.20	+	0.13
	worst radius	18.13	+	0.12
	worst area	1009.00	+	0.11
	mean area	838.10	+	0.06
	mean radius	16.50	+	0.05
	worst concavity	0.17	-	-0.05
Worst 2 of 2				
Predicted Probabilities: [benign: 1.0, malignant: 0.0]				
Predicted Value: benign				
Target Value: malignant				
Cross Entropy: 7.987				
Index ID: 135				

(continues on next page)

(continued from previous page)

SHAP Value	Feature Name	Feature Value	Contribution to Prediction
=====			
-0.03	mean concavity	0.05	-
-0.04	worst area	653.60	-
-0.05	worst concave points	0.09	-
-0.05	worst radius	14.49	-
-0.06	worst perimeter	92.04	-
-0.06	mean concave points	0.03	-

We use a custom metric ([hinge loss](#)) for selecting the best and worst predictions. See this example:

```
import numpy as np

def hinge_loss(y_true, y_pred_proba):
    probabilities = np.clip(y_pred_proba.iloc[:, 1], 0.001, 0.999)
    y_true[y_true == 0] = -1

    return np.clip(1 - y_true * np.log(probabilities / (1 - probabilities)), a_min=0,
a_max=None)

report = explain_predictions_best_worst(pipeline=pipeline, input_features=X, y_true=y,
include_shap_values=True, num_to_explain=5,
metric=hinge_loss)

print(report)
```

Changing Output Formats

Instead of getting the prediction explanations as text, you can get the report as a python dictionary or pandas dataframe. All you have to do is pass `output_format="dict"` or `output_format="dataframe"` to either `explain_prediction`, `explain_predictions`, or `explain_predictions_best_worst`.

Single prediction as a dictionary

```
[24]: import json
single_prediction_report = explain_predictions(pipeline=pipeline_binary, input_
↪features=X_holdout, indices_to_explain=[3],
                                             y=y_holdout, top_k_features=6, include_
↪shap_values=True,
                                             output_format="dict")
print(json.dumps(single_prediction_report, indent=2))

{
  "explanations": [
    {
      "explanations": [
        {
          "feature_names": [
            "worst concavity",
            "mean concavity",
            "worst area",
            "worst radius",
            "mean concave points",
            "worst perimeter"
          ],
          "feature_values": [
            0.1791,
            0.038,
            599.5,
            14.04,
            0.034,
            92.8
          ],
          "qualitative_explanation": [
            "-",
            "-",
            "-",
            "-",
            "-",
            "-"
          ],
          "quantitative_explanation": [
            -0.023008481104309524,
            -0.02621982146725469,
            -0.033821592020020774,
            -0.04666659740586632,
            -0.0541511910494414,
            -0.05523688273171911
          ],
          "drill_down": {},
          "class_name": "malignant",
          "expected_value": 0.3711208791208791
        }
      ]
    }
  ]
}
```


Single prediction as a dataframe

```
[25]: single_prediction_report = explain_predictions(pipeline=pipeline_binary, input_
      ↪ features=X_holdout,
                                     indices_to_explain=[3],
                                     y=y_holdout, top_k_features=6, include_
      ↪ shap_values=True,
                                     output_format="dataframe")
single_prediction_report
```

```
[25]:
```

	feature_names	feature_values	qualitative_explanation	\
0	worst concavity	0.1791		-
1	mean concavity	0.0380		-
2	worst area	599.5000		-
3	worst radius	14.0400		-
4	mean concave points	0.0340		-
5	worst perimeter	92.8000		-

	quantitative_explanation	class_name	prediction_number
0	-0.023008	malignant	0
1	-0.026220	malignant	0
2	-0.033822	malignant	0
3	-0.046667	malignant	0
4	-0.054151	malignant	0
5	-0.055237	malignant	0

Best and worst predictions as a dictionary

```
[26]: report = explain_predictions_best_worst(pipeline=pipeline_binary, input_features=X, y_
      ↪ true=y,
                                     num_to_explain=1, top_k_features=6,
                                     include_shap_values=True, output_format="dict"
      ↪ ")
print(json.dumps(report, indent=2))
```

```
{
  "explanations": [
    {
      "rank": {
        "prefix": "best",
        "index": 1
      },
      "predicted_values": {
        "probabilities": {
          "benign": 1.0,
          "malignant": 0.0
        },
        "predicted_value": "benign",
        "target_value": "benign",
        "error_name": "Cross Entropy",
        "error_value": 0.0001970443507070075,
        "index_id": 475
      },
      "explanations": [
        {
          "feature_names": [
```

(continues on next page)

(continued from previous page)

```

        "mean concavity",
        "worst area",
        "worst radius",
        "worst concave points",
        "worst perimeter",
        "mean concave points"
    ],
    "feature_values": [
        0.05835,
        605.8,
        14.09,
        0.09783,
        93.22,
        0.03078
    ],
    "qualitative_explanation": [
        "-",
        "-",
        "-",
        "-",
        "-",
        "-"
    ],
    "quantitative_explanation": [
        -0.028481050954786636,
        -0.03050522196002462,
        -0.042922079201003216,
        -0.04429366151003684,
        -0.05486784013962313,
        -0.05639460900233733
    ],
    "drill_down": {},
    "class_name": "malignant",
    "expected_value": 0.3711208791208791
    }
]
},
{
    "rank": {
        "prefix": "worst",
        "index": 1
    },
    "predicted_values": {
        "probabilities": {
            "benign": 1.0,
            "malignant": 0.0
        },
        "predicted_value": "benign",
        "target_value": "malignant",
        "error_name": "Cross Entropy",
        "error_value": 7.986911819330411,
        "index_id": 135
    },
    "explanations": [
        {
            "feature_names": [
                "mean concavity",

```

(continues on next page)

(continued from previous page)

```

        "worst area",
        "worst concave points",
        "worst radius",
        "worst perimeter",
        "mean concave points"
    ],
    "feature_values": [
        0.04711,
        653.6,
        0.09331,
        14.49,
        92.04,
        0.02704
    ],
    "qualitative_explanation": [
        "-",
        "-",
        "-",
        "-",
        "-",
        "-"
    ],
    "quantitative_explanation": [
        -0.029936744551331215,
        -0.03748357654576422,
        -0.04553126236476177,
        -0.0483274199182721,
        -0.06039220265366764,
        -0.060441902449258976
    ],
    "drill_down": {},
    "class_name": "malignant",
    "expected_value": 0.3711208791208791
    }
    ]
}

```

Best and worst predictions as a dataframe

```

[27]: report = explain_predictions_best_worst(pipeline=pipeline_binary, input_features=X_
↳holdout, y_true=y_holdout,
                                           num_to_explain=1, top_k_features=6,
                                           include_shap_values=True, output_format=
↳"dataframe")
report

```

```

[27]:
   feature_names  feature_values  qualitative_explanation  \
0    mean concavity          0.05928                    -
1      worst area          552.00000                    -
2  worst concave points          0.08411                    -
3      worst radius          13.57000                    -
4    mean concave points          0.03279                    -
5      worst perimeter          86.67000                    -

```

(continues on next page)

(continued from previous page)

6	mean concavity	0.04711	-
7	worst area	653.60000	-
8	worst concave points	0.09331	-
9	worst radius	14.49000	-
10	worst perimeter	92.04000	-
11	mean concave points	0.02704	-

	quantitative_explanation	class_name	label_benign_probability	\
0	-0.029022	malignant	1.0	
1	-0.034112	malignant	1.0	
2	-0.046896	malignant	1.0	
3	-0.046928	malignant	1.0	
4	-0.052902	malignant	1.0	
5	-0.064320	malignant	1.0	
6	-0.029937	malignant	1.0	
7	-0.037484	malignant	1.0	
8	-0.045531	malignant	1.0	
9	-0.048327	malignant	1.0	
10	-0.060392	malignant	1.0	
11	-0.060442	malignant	1.0	

	label_malignant_probability	predicted_value	target_value	error_name	\
0	0.0	benign	benign	Cross Entropy	
1	0.0	benign	benign	Cross Entropy	
2	0.0	benign	benign	Cross Entropy	
3	0.0	benign	benign	Cross Entropy	
4	0.0	benign	benign	Cross Entropy	
5	0.0	benign	benign	Cross Entropy	
6	0.0	benign	malignant	Cross Entropy	
7	0.0	benign	malignant	Cross Entropy	
8	0.0	benign	malignant	Cross Entropy	
9	0.0	benign	malignant	Cross Entropy	
10	0.0	benign	malignant	Cross Entropy	
11	0.0	benign	malignant	Cross Entropy	

	error_value	index_id	rank	prefix
0	0.000197	502	1	best
1	0.000197	502	1	best
2	0.000197	502	1	best
3	0.000197	502	1	best
4	0.000197	502	1	best
5	0.000197	502	1	best
6	7.986912	135	1	worst
7	7.986912	135	1	worst
8	7.986912	135	1	worst
9	7.986912	135	1	worst
10	7.986912	135	1	worst
11	7.986912	135	1	worst

Force Plots

Force plots can be generated to predict single or multiple rows for binary, multiclass and regression problem types. Here's an example of predicting a single row on a binary classification dataset. The force plots show the predictive power of each of the features in making the negative ("Class: 0") prediction and the positive ("Class: 1") prediction.

```
[28]: import shap

from evalml.model_understanding.force_plots import graph_force_plot

rows_to_explain = [0] # Should be a list of integer indices of the rows to explain.

results = graph_force_plot(pipeline_binary, rows_to_explain=rows_to_explain,
                           training_data=X_holdout, y=y_holdout)

for result in results:
    for cls in result:
        print("Class:", cls)
        display(result[cls]["plot"])

<IPython.core.display.HTML object>

Class: malignant

<shap.plots._force.AdditiveForceVisualizer at 0x7f9d665d8670>
```

Here's an example of a force plot explaining multiple predictions on a multiclass problem. These plots show the force plots for each row arranged as consecutive columns that can be ordered by the dropdown above. Clicking the column indicates which row explanation is underneath.

```
[29]: rows_to_explain = [0,1,2,3,4] # Should be a list of integer indices of the rows to_
    ↪explain.

results = graph_force_plot(pipeline_multi,
                           rows_to_explain=rows_to_explain,
                           training_data=X_multi, y=y_multi)

for idx, result in enumerate(results):
    print("Row:", idx)
    for cls in result:
        print("Class:", cls)
        display(result[cls]["plot"])

<IPython.core.display.HTML object>

Row: 0
Class: class_0

<shap.plots._force.AdditiveForceVisualizer at 0x7f9d66b94700>

Class: class_1

<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e130>

Class: class_2

<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e520>

Row: 1
Class: class_0

<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669efa0>
```

Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e040>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e070>
Row: 2
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669ed90>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e5e0>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669ef70>
Row: 3
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669ebb0>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6669e610>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6670efd0>
Row: 4
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6665ee0d0>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6665ee460>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f9d6665ee8e0>

4.7 Data Checks

EvalML provides data checks to help guide you in achieving the highest performing model. These utility functions help deal with problems such as overfitting, abnormal data, and missing data. These data checks can be found under `evalml/data_checks`. Below we will cover examples for each available data check in EvalML, as well as the `DefaultDataChecks` collection of data checks.

4.7.1 Missing Data

Missing data or rows with NaN values provide many challenges for machine learning pipelines. In the worst case, many algorithms simply will not run with missing data! EvalML pipelines contain imputation *components* to ensure that doesn't happen. Imputation works by approximating missing values with existing values. However, if a column contains a high number of missing values, a large percentage of the column would be approximated by a small percentage. This could potentially create a column without useful information for machine learning pipelines. By using `HighlyNullDataCheck`, EvalML will alert you to this potential problem by returning the columns that pass the missing values threshold.

```
[1]: import numpy as np
import pandas as pd

from evalml.data_checks import HighlyNullDataCheck

X = pd.DataFrame([[1, 2, 3],
                  [0, 4, np.nan],
                  [1, 4, np.nan],
                  [9, 4, np.nan],
                  [8, 6, np.nan]])

null_check = HighlyNullDataCheck(pct_null_col_threshold=0.8, pct_null_row_threshold=0.8)
results = null_check.validate(X)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])
```

Warning: Column '2' is 80.0% or more null

4.7.2 Abnormal Data

EvalML provides a few data checks to check for abnormal data:

- NoVarianceDataCheck
- ClassImbalanceDataCheck
- TargetLeakageDataCheck
- InvalidTargetDataCheck
- IDColumnsDataCheck
- OutliersDataCheck
- HighVarianceCVDDataCheck
- MulticollinearityDataCheck
- UniquenessDataCheck
- TargetDistributionDataCheck
- DateTimeFormatDataCheck

Zero Variance

Data with zero variance indicates that all values are identical. If a feature has zero variance, it is not likely to be a useful feature. Similarly, if the target has zero variance, there is likely something wrong. NoVarianceDataCheck checks if the target or any feature has only one unique value and alerts you to any such columns.

```
[2]: from evalml.data_checks import NoVarianceDataCheck
X = pd.DataFrame({"no var col": [0, 0, 0],
                  "good col": [0, 4, 1]})
y = pd.Series([1, 0, 1])
```

(continues on next page)

(continued from previous page)

```
no_variance_data_check = NoVarianceDataCheck()
results = no_variance_data_check.validate(X, y)
```

```
for message in results['warnings']:
    print("Warning:", message['message'])
```

```
for message in results['errors']:
    print("Error:", message['message'])
```

```
Error: no var col has 1 unique value.
```

Note that you can set NaN to count as an unique value, but `NoVarianceDataCheck` will still return a warning if there is only one unique non-NaN value in a given column.

```
[3]: from evalml.data_checks import NoVarianceDataCheck

X = pd.DataFrame({"no var col": [0, 0, 0],
                  "no var col with nan": [1, np.nan, 1],
                  "good col": [0, 4, 1]})
y = pd.Series([1, 0, 1])

no_variance_data_check = NoVarianceDataCheck(count_nan_as_value=True)
results = no_variance_data_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

Warning: no var col with nan has two unique values including nulls. Consider encoding
↳ the nulls for this column to be useful for machine learning.
Error: no var col has 1 unique value.
```

Class Imbalance

For classification problems, the distribution of examples across each class can vary. For small variations, this is normal and expected. However, when the number of examples for each class label is disproportionately biased or skewed towards a particular class (or classes), it can be difficult for machine learning models to predict well. In addition, having a low number of examples for a given class could mean that one or more of the CV folds generated for the training data could only have few or no examples from that class. This may cause the model to only predict the majority class and ultimately resulting in a poor-performant model.

`ClassImbalanceDataCheck` checks if the target labels are imbalanced beyond a specified threshold for a certain number of CV folds. It returns `DataCheckError` messages for any classes that have less samples than double the number of CV folds specified (since that indicates the likelihood of having at little to no samples of that class in a given fold), and `DataCheckWarning` messages for any classes that fall below the set threshold percentage.

```
[4]: from evalml.data_checks import ClassImbalanceDataCheck

X = pd.DataFrame([[1, 2, 0, 1],
                  [4, 1, 9, 0],
                  [4, 4, 8, 3],
                  [9, 2, 7, 1]])
y = pd.Series([0, 1, 1, 1, 1])
```

(continues on next page)

(continued from previous page)

```

class_imbalance_check = ClassImbalanceDataCheck(threshold=0.25, num_cv_folds=4)
results = class_imbalance_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

```

Warning: The following labels fall below 25% of the target: [0]
Warning: The following labels in the target have severe class imbalance because they
↳ fall under 25% of the target and have less than 100 samples: [0]
Error: The number of instances of these targets is less than 2 * the number of cross
↳ folds = 8 instances: [0, 1]

Target Leakage

Target leakage, also known as data leakage, can occur when you train your model on a dataset that includes information that should not be available at the time of prediction. This causes the model to score suspiciously well, but perform poorly in production. `TargetLeakageDataCheck` checks for features that could potentially be “leaking” information by calculating the Pearson correlation coefficient between each feature and the target to warn users if there are features are highly correlated with the target. Currently, only numerical features are considered.

```

[5]: from evalml.data_checks import TargetLeakageDataCheck
X = pd.DataFrame({'leak': [10, 42, 31, 51, 61],
                  'x': [42, 54, 12, 64, 12],
                  'y': [12, 5, 13, 74, 24]})
y = pd.Series([10, 42, 31, 51, 40])

target_leakage_check = TargetLeakageDataCheck(pct_corr_threshold=0.8)
results = target_leakage_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

```

Warning: Column 'leak' is 80.0% or more correlated with the target
Warning: Column 'x' is 80.0% or more correlated with the target
Warning: Column 'y' is 80.0% or more correlated with the target

Invalid Target Data

The `InvalidTargetDataCheck` checks if the target data contains any missing or invalid values. Specifically:

- if any of the target values are missing, a `DataCheckError` message is returned
- if the specified problem type is a binary classification problem but there is more or less than two unique values in the target, a `DataCheckError` message is returned
- if binary classification target classes are numeric values not equal to {0, 1}, a `DataCheckError` message is returned because it can cause unpredictable behavior when passed to pipelines

```
[6]: from evalml.data_checks import InvalidTargetDataCheck

X = pd.DataFrame({})
y = pd.Series([0, 1, None, None])

invalid_target_check = InvalidTargetDataCheck('binary', 'Log Loss Binary')
results = invalid_target_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])
```

Warning: Input target and features have different lengths
Warning: Input target and features have mismatched indices
Error: 2 row(s) (50.0%) of target values are null

ID Columns

ID columns in your dataset provide little to no benefit to a machine learning pipeline as the pipeline cannot extrapolate useful information from unique identifiers. Thus, `IDColumnsDataCheck` reminds you if these columns exists. In the given example, 'user_number' and 'id' columns are both identified as potentially being unique identifiers that should be removed.

```
[7]: from evalml.data_checks import IDColumnsDataCheck

X = pd.DataFrame([[0, 53, 6325, 5], [1, 90, 6325, 10], [2, 90, 18, 20]], columns=['user_
↪number', 'cost', 'revenue', 'id'])

id_col_check = IDColumnsDataCheck(id_threshold=0.9)
results = id_col_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])
```

Warning: Column 'id' is 90.0% or more likely to be an ID column
Warning: Column 'user_number' is 90.0% or more likely to be an ID column

Multicollinearity

The `MulticollinearityDataCheck` data check is used in to detect if are any set of features that are likely to be multicollinear. Multicollinear features affect the performance of a model, but more importantly, it may greatly impact model interpretation. EvalML uses mutual information to determine collinearity.

```
[8]: from evalml.data_checks import MulticollinearityDataCheck

y = pd.Series([1, 0, 2, 3, 4])
X = pd.DataFrame({'col_1': y,
                  'col_2': y * 3,
                  'col_3': ~y,
                  'col_4': y / 2,
```

(continues on next page)

(continued from previous page)

```

        'col_5': y + 1,
        'not_collinear': [0, 1, 0, 0, 0])

multi_check = MulticollinearityDataCheck(threshold=0.95)
results = multi_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

```

Warning: Columns are likely to be correlated: [('col_1', 'col_2'), ('col_1', 'col_3'),
 ↳ ('col_1', 'col_4'), ('col_1', 'col_5'), ('col_2', 'col_3'), ('col_2', 'col_4'), (
 ↳ 'col_2', 'col_5'), ('col_3', 'col_4'), ('col_3', 'col_5'), ('col_4', 'col_5')]

Uniqueness

The UniquenessDataCheck is used to detect columns with either too unique or not unique enough values. For regression type problems, the data is checked for a lower limit of uniqueness. For multiclass type problems, the data is checked for an upper limit.

```

[9]: import pandas as pd
    from evalml.data_checks import UniquenessDataCheck

X = pd.DataFrame({'most_unique': [float(x) for x in range(10)], # [0,1,2,3,4,5,6,7,8,
↳ 9]
                  'more_unique': [x % 5 for x in range(10)], # [0,1,2,3,4,0,1,2,3,4]
                  'unique': [x % 3 for x in range(10)], # [0,1,2,0,1,2,0,1,2,0]
                  'less_unique': [x % 2 for x in range(10)], # [0,1,0,1,0,1,0,1,0,1]
                  'not_unique': [float(1) for x in range(10)]}) # [1,1,1,1,1,1,1,1,1,1,
↳ 1]

uniqueness_check = UniquenessDataCheck(problem_type="regression",
                                       threshold=.5)
results = uniqueness_check.validate(X, y=None)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

```

Warning: Input columns (not_unique) for regression problem type are not unique enough.

Sparsity

The `SparsityDataCheck` is used to identify features that contain a sparsity of values.

```
[10]: from evalml.data_checks import SparsityDataCheck

X = pd.DataFrame({'most_sparse': [float(x) for x in range(10)], # [0,1,2,3,4,5,6,7,8,
    ↪ 9]
                  'more_sparse': [x % 5 for x in range(10)], # [0,1,2,3,4,0,1,2,3,
    ↪ 4]
                  'sparse': [x % 3 for x in range(10)], # [0,1,2,0,1,2,0,1,2,
    ↪ 0]
                  'less_sparse': [x % 2 for x in range(10)], # [0,1,0,1,0,1,0,1,0,
    ↪ 1]
                  'not_sparse': [float(1) for x in range(10)]}) # [1,1,1,1,1,1,1,1,1,
    ↪ 1]

sparsity_check = SparsityDataCheck(problem_type="multiclass",
                                   threshold=.4,
                                   unique_count_threshold=3)
results = sparsity_check.validate(X,y=None)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

Warning: Input columns (most_sparse) for multiclass problem type are too sparse.
Warning: Input columns (more_sparse) for multiclass problem type are too sparse.
Warning: Input columns (sparse) for multiclass problem type are too sparse.
```

Outliers

Outliers are observations that differ significantly from other observations in the same sample. Many machine learning pipelines suffer in performance if outliers are not dropped from the training set as they are not representative of the data. `OutliersDataCheck()` uses IQR to notify you if a sample can be considered an outlier.

Below we generate a random dataset with some outliers.

```
[11]: data = np.tile(np.arange(10) * 0.01, (100, 10))
X = pd.DataFrame(data=data)

# generate some outliers in columns 3, 25, 55, and 72
X.iloc[0, 3] = -10000
X.iloc[3, 25] = 10000
X.iloc[5, 55] = 10000
X.iloc[10, 72] = -10000
```

We then utilize `OutliersDataCheck()` to rediscover these outliers.

```
[12]: from evalml.data_checks import OutliersDataCheck

outliers_check = OutliersDataCheck()
results = outliers_check.validate(X, y)
```

(continues on next page)

(continued from previous page)

```
for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])
```

```
Warning: Column(s) '3', '25', '55', '72' are likely to have outlier data.
```

Target Distribution

Target data can come in a variety of distributions, such as Gaussian or Lognormal. When we work with machine learning models, we feed data into an estimator that learns from the training data provided. Sometimes the data can be significantly spread out with a long tail or outliers, which could lead to a lognormal distribution. This can cause machine learning model performance to suffer.

To help the estimators better understand the underlying relationships in the data between the features and the target, we can use the `TargetDistributionDataCheck` to identify such a distribution. If you use `AutoML.search` to try and find the best pipeline, it will automatically run this data check for you and, if a lognormal distribution is detected, will add a `LogTransformer` to your pipeline to help the model performance!

```
[13]: from scipy.stats import lognorm
from evalml.data_checks import TargetDistributionDataCheck

data = np.tile(np.arange(10) * 0.01, (100, 10))
X = pd.DataFrame(data=data)
y = pd.Series(lognorm.rvs(s=0.4, loc=1, scale=1, size=100))

target_dist_check = TargetDistributionDataCheck()
results = target_dist_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

Warning: Target may have a lognormal distribution.
```

Datetime Format

Datetime information is a necessary component of time series problems, but sometimes the data we deal with may contain flaws that make it impossible for time series models to work with them. For example, in order to identify a frequency in the datetime information there has to be equal interval spacing between data points i.e. January 1, 2021, January 3, 2021, January 5, 2021, ...etc which are separated by two days. If instead there are random jumps in the datetime data i.e. January 1, 2021, January 3, 2021, January 12, 2021, then a frequency can't be inferred. Another common issue with time series models are that they can't handle datetime information that isn't properly sorted. Datetime values that aren't monotonically increasing (sorted in ascending order) will encounter this issue and their frequency cannot be inferred.

To make it easy to verify that the datetime column you're working with is properly spaced and sorted, we can leverage the `DatetimeFormatDataCheck`. When initializing the data check, pass in the name of the column that contains your datetime information (or pass in "index" if it's found in either your X or y indices).

```
[14]: from evalml.data_checks import DateTimeFormatDataCheck

X = pd.DataFrame(pd.date_range("January 1, 2021", periods=8, freq='2D'), columns=[
    ↪ "dates"])
y = pd.Series([1, 2, 4, 2, 1, 2, 3, 1])

# Replaces the last entry with January 16th instead of January 15th
# so that the data is no longer evenly spaced.
X.iloc[7] = "January 16, 2021"

datetime_format_check = DateTimeFormatDataCheck(datetime_column="dates")
results = datetime_format_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

print("-----")

# Reverses the order of the index datetime values to be decreasing.
X = X[::-1]
results = datetime_format_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])

for message in results['errors']:
    print("Error:", message['message'])

Error: No frequency could be detected in dates, possibly due to uneven intervals.
-----
Error: No frequency could be detected in dates, possibly due to uneven intervals.
Error: Datetime values must be sorted in ascending order.
```

4.7.3 Data Check Messages

Each data check's `validate` method returns a list of `DataCheckMessage` objects indicating warnings or errors found; warnings are stored as a `DataCheckWarning` object ([API reference](#)) and errors are stored as a `DataCheckError` object ([API reference](#)). You can filter the messages returned by a data check by checking for the type of message returned. Below, `NoVarianceDataCheck` returns a list containing a `DataCheckWarning` and a `DataCheckError` message. We can determine which is which by checking the type of each message.

```
[15]: from evalml.data_checks import NoVarianceDataCheck, DataCheckError, DataCheckWarning

X = pd.DataFrame({"no var col": [0, 0, 0],
                  "no var col with nan": [1, np.nan, 1],
                  "good col": [0, 4, 1]})
y = pd.Series([1, 0, 1])

no_variance_data_check = NoVarianceDataCheck(count_nan_as_value=True)
results = no_variance_data_check.validate(X, y)

for message in results['warnings']:
    print("Warning:", message['message'])
```

(continues on next page)

(continued from previous page)

```
for message in results['errors']:
    print("Error:", message['message'])
```

Warning: no var col with nan has two unique values including nulls. Consider encoding
 ↳ the nulls for this column to be useful for machine learning.
 Error: no var col has 1 unique value.

4.7.4 Writing Your Own Data Check

If you would prefer to write your own data check, you can do so by extending the `DataCheck` class and implementing the `validate(self, X, y)` class method. Below, we've created a new `DataCheck`, `ZeroVarianceDataCheck`, which is similar to `NoVarianceDataCheck` defined in EvalML. The `validate(self, X, y)` method should return a dictionary with 'warnings' and 'errors' as keys mapping to list of warnings and errors, respectively.

```
[16]: from evalml.data_checks import DataCheck

class ZeroVarianceDataCheck(DataCheck):
    def validate(self, X, y):
        messages = {'warnings': [], 'errors': []}
        if not isinstance(X, pd.DataFrame):
            X = pd.DataFrame(X)
        warning_msg = "Column '{}' has zero variance"
        messages['warnings'].extend([DataCheckError(warning_msg.format(column), self.
    ↳ name) for column in X.columns if len(X[column].unique()) == 1])
```

4.7.5 Defining Collections of Data Checks

For convenience, EvalML provides a `DataChecks` class to represent a collection of data checks. We will go over `DefaultDataChecks` ([API reference](#)), a collection defined to check for some of the most common data issues.

Default Data Checks

`DefaultDataChecks` is a collection of data checks defined to check for some of the most common data issues. They include:

- `HighlyNullDataCheck`
- `IDColumnsDataCheck`
- `TargetLeakageDataCheck`
- `InvalidTargetDataCheck`
- `TargetDistributionDataCheck` (for regression problem types)
- `ClassImbalanceDataCheck` (for classification problem types)
- `NoVarianceDataCheck`
- `DateTimeNaNDataCheck`
- `NaturalLanguageNaNDataCheck`
- `DateTimeFormatDataCheck` (for time series problem types)

4.7.6 Writing Your Own Collection of Data Checks

If you would prefer to create your own collection of data checks, you could either write your own data checks class by extending the `DataChecks` class and setting the `self.data_checks` attribute to the list of `DataCheck` classes or objects, or you could pass that list of data checks to the constructor of the `DataChecks` class. Below, we create two identical collections of data checks using the two different methods.

```
[17]: # Create a subclass of `DataChecks`
from evalml.data_checks import DataChecks, HighlyNullDataCheck,
↳InvalidTargetDataCheck, NoVarianceDataCheck, ClassImbalanceDataCheck,
↳TargetLeakageDataCheck
from evalml.problem_types import ProblemTypes, handle_problem_types

class MyCustomDataChecks(DataChecks):

    data_checks = [HighlyNullDataCheck, InvalidTargetDataCheck, NoVarianceDataCheck,
↳TargetLeakageDataCheck]

    def __init__(self, problem_type, objective):
        """
        A collection of basic data checks.
        Arguments:
            problem_type (str): The problem type that is being validated. Can be
↳regression, binary, or multiclass.
        """
        if handle_problem_types(problem_type) == ProblemTypes.REGRESSION:
            super().__init__(self.data_checks,
                             data_check_params={"InvalidTargetDataCheck": {"problem_
↳type": problem_type,
                                                                           "objective
↳": objective}})
        else:
            super().__init__(self.data_checks + [ClassImbalanceDataCheck],
                             data_check_params={"InvalidTargetDataCheck": {"problem_
↳type": problem_type,
                                                                           "objective
↳": objective}})

custom_data_checks = MyCustomDataChecks(problem_type=ProblemTypes.REGRESSION,
↳objective="R2")
for data_check in custom_data_checks.data_checks:
    print(data_check.name)
```

```
HighlyNullDataCheck
InvalidTargetDataCheck
NoVarianceDataCheck
TargetLeakageDataCheck
```

```
[18]: # Pass list of data checks to the `data_checks` parameter of DataChecks
same_custom_data_checks = DataChecks(data_checks=[HighlyNullDataCheck,
↳InvalidTargetDataCheck, NoVarianceDataCheck, TargetLeakageDataCheck],
                                     data_check_params={"InvalidTargetDataCheck": {
↳"problem_type": ProblemTypes.REGRESSION,
                                     "objective": "R2"}})
for data_check in same_custom_data_checks.data_checks:
    print(data_check.name)
```



```
HighlyNullDataCheck
InvalidTargetDataCheck
NoVarianceDataCheck
TargetLeakageDataCheck
```

4.8 Understanding Data Actions

EvalML streamlines the creation and implementation of machine learning models for tabular data. One of the many features it offers is [data checks](#), which are geared towards determining the health of the data before we train a model on it. These data checks have associated actions with them and will be shown in this notebook. In our default data checks, we have the following checks:

- `HighlyNullDataCheck`: Checks whether the rows or columns are highly null
- `IDColumnsDataCheck`: Checks for columns that could be ID columns
- `TargetLeakageDataCheck`: Checks if any of the input features have high association with the targets
- `InvalidTargetDataCheck`: Checks if there are null or other invalid values in the target
- `NoVarianceDataCheck`: Checks if either the target or any features have no variance
- `NaturalLanguageNaNDataCheck`: Checks if any natural language columns have missing data
- `DateTimeNaNDataCheck`: Checks if any datetime columns have missing data

EvalML has additional data checks that can be seen [here](#), with usage examples [here](#). Below, we will walk through usage of EvalML's default data checks and actions.

First, we import the necessary requirements to demonstrate these checks.

```
[1]: import woodwork as ww
import pandas as pd
from evalml import AutoMLSearch
from evalml.demos import load_fraud
from evalml.preprocessing import split_data
```

Let's look at the input feature data. EvalML uses the [Woodwork](#) library to represent this data. The demo data that EvalML returns is a Woodwork DataTable and DataColumn.

```
[2]: X, y = load_fraud(n_rows=1500)
X
```

```

      Number of Features
Boolean                1
Categorical            6
Numeric                5

Number of training examples: 1500
Targets
False      86.60%
True       13.40%
Name: fraud, dtype: object
```

```
[2]:      card_id  store_id      datetime  amount  currency  \
id
0      32261      8516 2019-01-01 00:12:26   24900      CUC
1      16434      8516 2019-01-01 09:42:03   15789      MYR
```

(continues on next page)

(continued from previous page)

```

2      23468      8516 2019-04-17 08:17:01      1883      AUD
3      14364      8516 2019-01-30 11:54:30      82120     KRW
4      29407      8516 2019-05-01 17:59:36      25745     MUR
...      ...      ...      ...      ...      ...
1495    9389      9389 2019-06-05 04:09:15      70510     XDR
1496    21867     7301 2019-03-02 18:32:43      75234     CNY
1497    10016     7301 2019-02-05 06:14:25      61331     LKR
1498    15971     7301 2019-01-03 13:19:19      77846     GTQ
1499    23638     7301 2019-04-26 17:20:41      76886     USD

      customer_present expiration_date      provider      lat \
id
0      True      08/24      Mastercard 38.58894
1      False     11/21      Discover 38.58894
2      False     09/27      Discover 38.58894
3      True      09/20      JCB 16 digit 38.58894
4      True      09/22      American Express 38.58894
...      ...      ...      ...      ...
1495    False     06/20      VISA 16 digit 9.53500
1496    True      02/23      Diners Club / Carte Blanche 37.49073
1497    True      02/27      VISA 13 digit 37.49073
1498    True      01/24      JCB 15 digit 37.49073
1499    True      02/29      American Express 37.49073

      lng      region country
id
0      -89.99038 Fairview Heights      US
1      -89.99038 Fairview Heights      US
2      -89.99038 Fairview Heights      US
3      -89.99038 Fairview Heights      US
4      -89.99038 Fairview Heights      US
...      ...      ...      ...
1495 -13.68778      Camayenne      GN
1496  -2.77259      Baza      ES
1497  -2.77259      Baza      ES
1498  -2.77259      Baza      ES
1499  -2.77259      Baza      ES

[1500 rows x 12 columns]
```

4.8.1 Adding noise and unclean data

This data is already clean and compatible with EvalML's `AutoMLSearch`. In order to demonstrate EvalML default data checks, we will add the following:

- A column of mostly null values (<0.5% non-null)
- A column with low/no variance
- A row of null values
- A missing target value

We will add the first two columns to the whole dataset and we will only add the last two to the training data. Note: these only represent some of the scenarios that EvalML default data checks can catch.

```
[3]: # add a column with no variance in the data
X['no_variance'] = [1 for _ in range(X.shape[0])]

# add a column with >99.5% null values
X['mostly_nulls'] = [None] * (X.shape[0] - 5) + [i for i in range(5)]

# since we changed the data, let's reinitialize the woodwork datatable
X.ww.init()
# let's split some training and validation data
X_train, X_valid, y_train, y_valid = split_data(X, y, problem_type='binary')
```

```
[4]: # let's copy the datetime at row 1 for future use
date = X_train.iloc[1]['datetime']

# make row 1 all nan values
X_train.iloc[1] = [None] * X_train.shape[1]

# make one of the target values null
y_train[990] = None

X_train.ww.init()
y_train = ww.init_series(y_train)
# Let's take another look at the new X_train data
X_train
```

```
[4]:
```

	card_id	store_id		datetime	amount	currency	\
id							
872	15492.0	2868.0	2019-08-03	02:50:04	80719.0	HNL	
1477	NaN	NaN		NaT	NaN	NaN	
158	22440.0	6813.0	2019-07-12	11:07:25	1849.0	SEK	
808	8096.0	8096.0	2019-06-11	21:33:36	41358.0	MOP	
336	33270.0	1529.0	2019-03-23	21:44:00	32594.0	CUC	
...	
339	8484.0	5358.0	2019-01-10	07:47:28	89503.0	GMD	
1383	17565.0	3929.0	2019-01-15	01:11:02	14264.0	DKK	
893	108.0	44.0	2019-05-17	00:53:39	93218.0	SLL	
385	29983.0	152.0	2019-06-09	06:50:29	41105.0	RWF	
1074	26197.0	4927.0	2019-05-22	15:57:27	50481.0	MNT	

	customer_present	expiration_date		provider	lat	lng	\
id							
872	True	08/27		American Express	5.47090	100.24529	
1477	NaN	NaN		NaN	NaN	NaN	
158	True	09/20		American Express	26.26490	81.54855	
808	True	04/29		VISA 13 digit	59.37722	28.19028	
336	False	04/22		Mastercard	51.39323	0.47713	
...	
339	False	11/24		Maestro	47.30997	8.52462	
1383	True	06/20		VISA 13 digit	50.72043	11.34046	
893	True	12/24		JCB 16 digit	15.72892	120.57224	
385	False	07/20		JCB 16 digit	-6.80000	39.25000	
1074	False	05/26		JCB 15 digit	41.00510	-73.78458	

	region	country	no_variance	mostly_nulls
id				
872	Batu Feringgi	MY	1.0	NaN
1477	NaN	NaN	NaN	NaN

(continues on next page)

(continued from previous page)

```

158      Jais      IN      1.0      NaN
808      Narva     EE      1.0      NaN
336      Strood    GB      1.0      NaN
...      ...      ...      ...      ...
339      Adliswil  CH      1.0      NaN
1383     Rudolstadt DE      1.0      NaN
893      Burgos   PH      1.0      NaN
385      Magomeni TZ      1.0      NaN
1074     Scarsdale US      1.0      NaN

[1200 rows x 14 columns]

```

If we call `AutoMLSearch.search()` on this data, the search will fail due to the columns and issues we've added above. Note: we use a `try/except` here to catch the resulting `ValueError` that `AutoMLSearch` raises.

```

[5]: automl = AutoMLSearch(X_train=X_train, y_train=y_train, problem_type='binary')
try:
    automl.search()
except ValueError as e:
    # to make the error message more distinct
    print("=" * 80, "\n")
    print("Search errored out! Message received is: {}".format(e))
    print("=" * 80, "\n")

=====

Search errored out! Message received is: Input contains NaN, infinity or a value too
↳large for dtype('float64').

=====

```

We can use the `search_iterative()` function provided in `EvalML` to determine what potential health issues our data has. We can see that this `search_iterative` function is a public method available through `evalml.automl` and is different from the `search` function of the `AutoMLSearch` class in `EvalML`. This `search_iterative()` function allows us to run the default data checks on the data, and, if there are no errors, automatically runs `AutoMLSearch.search()`.

```

[6]: from evalml.automl import search_iterative
results = search_iterative(X_train, y_train, problem_type='binary')
results

[6]: (None,
      {'warnings': [{'message': '1 out of 1200 rows are more than 95.0% null',
                              'data_check_name': 'HighlyNullDataCheck',
                              'level': 'warning',
                              'code': 'HIGHLY_NULL_ROWS',
                              'details': {'pct_null_cols': id
                                           1477      1.0
                                           dtype: float64}},
                    {'message': "Column 'mostly_nulls' is 95.0% or more null",
                              'data_check_name': 'HighlyNullDataCheck',
                              'level': 'warning',
                              'code': 'HIGHLY_NULL_COLS',
                              'details': {'column': 'mostly_nulls',
                                           'pct_null_rows': 0.9966666666666667}}]},
      {'errors': [{'message': '1 row(s) (0.083333333333333334%) of target values are null',
                    'data_check_name': 'InvalidTargetDataCheck',

```

(continues on next page)

(continued from previous page)

```

    'level': 'error',
    'code': 'TARGET_HAS_NULL',
    'details': {'num_null_rows': 1, 'pct_null_rows': 0.08333333333333334}},
    {'message': 'no_variance has 1 unique value.',
     'data_check_name': 'NoVarianceDataCheck',
     'level': 'error',
     'code': 'NO_VARIANCE',
     'details': {'column': 'no_variance'}},
    {'message': 'Input datetime column(s) (datetime) contains NaN values. Please_
↳impute NaN values or drop these rows or columns.',
     'data_check_name': 'DateTimeNaNDataCheck',
     'level': 'error',
     'code': 'DATETIME_HAS_NAN',
     'details': {'columns': 'datetime'}}],
    'actions': [{
        'code': 'DROP_ROWS', 'metadata': {'rows': [1477]}},
        {'code': 'DROP_COL', 'metadata': {'column': 'mostly_nulls'}},
        {'code': 'IMPUTE_COL',
         'metadata': {'column': None,
                     'is_target': True,
                     'impute_strategy': 'most_frequent'}},
        {'code': 'DROP_COL', 'metadata': {'column': 'no_variance'}}}]

```

The return value of the `search_iterative` function above is a tuple. The first element is the `AutoMLSearch` object if it runs (and `None` otherwise), and the second element is a dictionary of potential warnings and errors that the default data checks find on the passed-in `X` and `y` data. In this dictionary, warnings are suggestions that the datachecks give that can be useful to address to make the search better but will not break `AutoMLSearch`. On the flip side, errors will break `AutoMLSearch` and need to be addressed by the user.

4.8.2 Addressing DataCheck errors

We will show that we can address errors to allow `AutoMLSearch` to run. However, ignoring warnings will come at the expense of performance.

We can print out the errors first to make it easier to read, and then we'll create new features and targets from the original training data.

```

[7]: results[1]['errors']
[7]: [{'message': '1 row(s) (0.08333333333333334%) of target values are null',
      'data_check_name': 'InvalidTargetDataCheck',
      'level': 'error',
      'code': 'TARGET_HAS_NULL',
      'details': {'num_null_rows': 1, 'pct_null_rows': 0.08333333333333334}},
     {'message': 'no_variance has 1 unique value.',
      'data_check_name': 'NoVarianceDataCheck',
      'level': 'error',
      'code': 'NO_VARIANCE',
      'details': {'column': 'no_variance'}},
     {'message': 'Input datetime column(s) (datetime) contains NaN values. Please impute_
↳NaN values or drop these rows or columns.',
      'data_check_name': 'DateTimeNaNDataCheck',
      'level': 'error',
      'code': 'DATETIME_HAS_NAN',
      'details': {'columns': 'datetime'}}]

```

```
[8]: # copy the DataTables to new variables
X_train_no_errors = X_train.copy()
y_train_no_errors = y_train.copy()

# We address the errors by looking at the resulting dictionary errors listed

# first, let's address the `TARGET_HAS_NULL` error
y_train_no_errors.fillna(False, inplace=True)

# here, we address the `NO_VARIANCE` error
X_train_no_errors.drop("no_variance", axis=1, inplace=True)

# lastly, we address the `DATETIME_HAS_NAN` error with the date we had saved earlier
X_train_no_errors.iloc[1, 2] = date

# let's reinitialize the Woodwork DataTable
X_train_no_errors.ww.init()
X_train_no_errors.head()
```

```
[8]:
```

	card_id	store_id		datetime	amount	currency	\
id							
872	15492.0	2868.0	2019-08-03	02:50:04	80719.0	HNL	
1477	NaN	NaN	2019-08-05	21:05:57	NaN	NaN	
158	22440.0	6813.0	2019-07-12	11:07:25	1849.0	SEK	
808	8096.0	8096.0	2019-06-11	21:33:36	41358.0	MOP	
336	33270.0	1529.0	2019-03-23	21:44:00	32594.0	CUC	

	customer_present	expiration_date		provider	lat	lng	\
id							
872	True	08/27		American Express	5.47090	100.24529	
1477	NaN	NaN		NaN	NaN	NaN	
158	True	09/20		American Express	26.26490	81.54855	
808	True	04/29		VISA 13 digit	59.37722	28.19028	
336	False	04/22		Mastercard	51.39323	0.47713	

	region	country	mostly_nulls
id			
872	Batu Feringgi	MY	NaN
1477	NaN	NaN	NaN
158	Jais	IN	NaN
808	Narva	EE	NaN
336	Strood	GB	NaN

We can now run search on `X_train_no_errors` and `y_train_no_errors`. Note that the search here doesn't fail since we addressed the errors, but there will still exist warnings in the returned tuple. This search allows the `mostly_nulls` column to remain in the features during search.

```
[9]: results_no_errors = search_iterative(X_train_no_errors, y_train_no_errors, problem_
    ↪type='binary')
results_no_errors
```

High coefficient of variation (cv >= 0.5) within cross validation scores.
 Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One_

↪Hot Encoder + Oversampler may not perform as estimated on unseen data.

```
[9]: (<evalml.automl.automl_search.AutoMLSearch at 0x7f2a66e48a60>,
    {'warnings': [{'message': "Column 'mostly_nulls' is 95.0% or more null",
        'data_check_name': 'HighlyNullDataCheck',
        'level': 'warning',
```

(continues on next page)

(continued from previous page)

```
'code': 'HIGHLY_NULL_COLS',
'details': {'column': 'mostly_nulls',
'pct_null_rows': 0.9966666666666667}},
'errors': [],
'actions': [{'code': 'DROP_COL', 'metadata': {'column': 'mostly_nulls'}}])
```

4.8.3 Addressing all warnings and errors

We can look at the actions key of the dictionary in order to see how we can fix and clean all of the data. This will help us clean both the warnings and errors from the data and provide us with a better model.

```
[10]: results[1]['actions']
[10]: [{'code': 'DROP_ROWS', 'metadata': {'rows': [1477]}},
{'code': 'DROP_COL', 'metadata': {'column': 'mostly_nulls'}},
{'code': 'IMPUTE_COL',
'metadata': {'column': None,
'is_target': True,
'impute_strategy': 'most_frequent'}},
{'code': 'DROP_COL', 'metadata': {'column': 'no_variance'}}]
```

We note that there are four action tasks that we can take to clean the data. Three of the tasks ask us to drop a row or column in the features, while one task asks us to impute the target value.

```
[11]: # The first action states to drop the row given by the action code
X_train.drop(1477, axis=0, inplace=True)
# we must also drop this for y since we are removing its associated feature input
y_train.drop(index=1477, inplace=True)

print("The new length of X_train is {} and y_train is {}".format(len(X_train), len(y_
→train)))

The new length of X_train is 1199 and y_train is 1199
```

```
[12]: # Remove the 'mostly_nulls' column from X_train, which is the second action item
X_train.drop('mostly_nulls', axis=1, inplace=True)
X_train.head()
```

```
[12]:
```

	card_id	store_id		datetime	amount	currency	\
id							
872	15492.0	2868.0	2019-08-03	02:50:04	80719.0	HNL	
158	22440.0	6813.0	2019-07-12	11:07:25	1849.0	SEK	
808	8096.0	8096.0	2019-06-11	21:33:36	41358.0	MOP	
336	33270.0	1529.0	2019-03-23	21:44:00	32594.0	CUC	
1262	22829.0	8390.0	2019-01-23	04:39:29	86149.0	SCR	

	customer_present	expiration_date		provider	lat	lng	\
id							
872	True	08/27		American Express	5.47090	100.24529	
158	True	09/20		American Express	26.26490	81.54855	
808	True	04/29		VISA 13 digit	59.37722	28.19028	
336	False	04/22		Mastercard	51.39323	0.47713	
1262	True	03/25		Maestro	48.84490	17.22635	

	region	country	no_variance
id			

(continues on next page)

(continued from previous page)

872	Batu Feringgi	MY	1.0
158	Jais	IN	1.0
808	Narva	EE	1.0
336	Strood	GB	1.0
1262	Skalica	SK	1.0

```
[13]: # Address the null in targets, which is the third action item
y_train.fillna(False, inplace=True)
y_train.isna().any()
```

```
[13]: False
```

```
[14]: # Finally, we can drop the 'no_variance' column, which is the final action item
X_train.drop('no_variance', axis=1, inplace=True)
X_train.head()
```

```
[14]:
```

	card_id	store_id		datetime	amount	currency	\
id							
872	15492.0	2868.0	2019-08-03	02:50:04	80719.0	HNL	
158	22440.0	6813.0	2019-07-12	11:07:25	1849.0	SEK	
808	8096.0	8096.0	2019-06-11	21:33:36	41358.0	MOP	
336	33270.0	1529.0	2019-03-23	21:44:00	32594.0	CUC	
1262	22829.0	8390.0	2019-01-23	04:39:29	86149.0	SCR	

	customer_present	expiration_date		provider	lat	lng	\
id							
872	True	08/27		American Express	5.47090	100.24529	
158	True	09/20		American Express	26.26490	81.54855	
808	True	04/29		VISA 13 digit	59.37722	28.19028	
336	False	04/22		Mastercard	51.39323	0.47713	
1262	True	03/25		Maestro	48.84490	17.22635	

	region	country
id		
872	Batu Feringgi	MY
158	Jais	IN
808	Narva	EE
336	Strood	GB
1262	Skalica	SK

```
[15]: # let's reinitialize the dataframe using Woodwork and try the search again
X_train.ww.init()
results_cleaned = search_iterative(X_train, y_train, problem_type='binary')
```

Note that this time, we do get an AutoMLSearch object returned to us, as well as an empty dictionary of warnings and errors. We can use the AutoMLSearch object as needed, and we can see that the resulting warning dictionary is empty.

```
[16]: aml = results_cleaned[0]
aml.rankings
```

```
[16]:
```

	id		pipeline_name	search_order	\
0	6		XGBoost Classifier w/ Imputer + DateTime Featu...	6	
1	3		Random Forest Classifier w/ Imputer + DateTime...	3	
2	4		LightGBM Classifier w/ Imputer + DateTime Feat...	4	
3	7		Extra Trees Classifier w/ Imputer + DateTime F...	7	
4	1		Elastic Net Classifier w/ Imputer + DateTime F...	1	

(continues on next page)

(continued from previous page)

5	5	Logistic Regression Classifier w/ Imputer + Da...	5
6	8	CatBoost Classifier w/ Imputer + DateTime Feat...	8
7	2	Decision Tree Classifier w/ Imputer + DateTime...	2
8	0	Mode Baseline Binary Classification Pipeline	0

	mean_cv_score	standard_deviation_cv_score	validation_score \
0	0.245210	0.033784	0.283381
1	0.261529	0.015703	0.279336
2	0.342549	0.068008	0.407373
3	0.350142	0.007130	0.358375
4	0.392373	0.014820	0.404130
5	0.394584	0.014575	0.405650
6	0.546942	0.001789	0.548985
7	0.632121	0.156515	0.665526
8	4.637776	0.043230	4.662735

	percent_better_than_baseline	high_variance_cv \
0	94.712762	False
1	94.360903	False
2	92.613932	False
3	92.450223	False
4	91.539639	False
5	91.491950	False
6	88.206800	False
7	86.370173	False
8	0.000000	False

	parameters
0	{'Imputer': {'categorical_impute_strategy': 'm...
1	{'Imputer': {'categorical_impute_strategy': 'm...
2	{'Imputer': {'categorical_impute_strategy': 'm...
3	{'Imputer': {'categorical_impute_strategy': 'm...
4	{'Imputer': {'categorical_impute_strategy': 'm...
5	{'Imputer': {'categorical_impute_strategy': 'm...
6	{'Imputer': {'categorical_impute_strategy': 'm...
7	{'Imputer': {'categorical_impute_strategy': 'm...
8	{'Baseline Classifier': {'strategy': 'mode'}}

```
[17]: data_check_results = results_cleaned[1]
      data_check_results
```

```
[17]: {'warnings': [], 'errors': [], 'actions': []}
```

4.8.4 Comparing removing only errors versus removing both warnings and errors

Let's see the differences in model performance when we remove only errors versus remove both warnings and errors. To do this, we compare the performance of the best pipelines on the validation data. Remember that in the search where we only address errors, we still have the `mostly_nulls` column present in the data, so we leave that column in the validation data for its respective search. We drop the other `no_variance` column from both searches.

Additionally, we do some logical type setting since we had added additional noise to just the training data. This allows the data to be of the same types in both training and validation.

```
[18]: # drop the no_variance column
      X_valid.drop("no_variance", axis=1, inplace=True)
```

(continues on next page)

(continued from previous page)

```
# logical type management
X_valid.ww.init(logical_types={"customer_present": "Categorical"})
y_valid = ww.init_series(y_valid, logical_type="Categorical")

best_pipeline_no_errors = results_no_errors[0].best_pipeline
print("Only dropping errors:", best_pipeline_no_errors.score(X_valid, y_valid, ["Log_
↳ Loss Binary"]), "\n")

# drop the mostly_nulls column and reinitialize the DataTable
X_valid.drop("mostly_nulls", axis=1, inplace=True)
X_valid.ww.init()

best_pipeline_clean = results_cleaned[0].best_pipeline
print("Addressing all actions:", best_pipeline_clean.score(X_valid, y_valid, ["Log_
↳ Loss Binary"]), "\n")

Only dropping errors: OrderedDict([('Log Loss Binary', 0.2252137942060951)])

Addressing all actions: OrderedDict([('Log Loss Binary', 0.2282781904970938)])
```

We can compare the differences in model performance when we address all action items (warnings and errors) in comparison to when we only address errors. While it isn't guaranteed that addressing all actions will always have better performance, we do recommend doing so since we only raise these issues when we believe the features have problems that could negatively impact or not benefit the search.

In the future, we aim to provide a helper function to allow users to quickly clean the data by taking in the list of actions and creating an appropriate pipeline of transformers to alter the data.

4.9 Utilities

4.9.1 Configuring Logging

EvalML uses [the standard Python logging package](#). Default logging behavior prints WARNING level logs and above (ERROR and CRITICAL) to stdout. To configure different behavior, please refer to the Python logging documentation.

To see up-to-date feedback as `AutoMLSearch` runs, use the argument `verbose=True` when instantiating the object. This will temporarily set up a logging object to print INFO level logs and above to stdout, as well as display a graph of the best score over pipeline iterations.

4.9.2 System Information

EvalML provides a command-line interface (CLI) tool prints the version of EvalML and core dependencies installed, as well as some basic system information. To use this tool, just run `evalml info` in your shell or terminal. This could be useful for debugging purposes or tracking down any version-related issues.

```
[1]: !evalml info

EvalML version: 0.32.1
EvalML installation directory: /home/docs/checkouts/readthedocs.org/user_builds/
↳ feature-labs-inc-evalml/envs/v0.32.1/lib/python3.8/site-packages/evalml
```

(continues on next page)

(continued from previous page)

SYSTEM INFO

```
python: 3.8.6.final.0
python-bits: 64
OS: Linux
OS-release: 5.4.0-1035-aws
machine: x86_64
processor: x86_64
byteorder: little
LC_ALL: None
LANG: C.UTF-8
LOCALE: en_US.UTF-8
# of CPUs: 2
Available memory: 6.0G
```

INSTALLED VERSIONS

```
zict: 2.0.0
xgboost: 1.4.2
wrapt: 1.12.1
woodwork: 0.7.1
widgetsnbextension: 3.5.1
wheel: 0.37.0
webencodings: 0.5.1
wcwidth: 0.2.5
urllib3: 1.26.6
unidecode: 1.3.1
traitlets: 5.1.0
tqdm: 4.62.2
tornado: 6.1
toolz: 0.11.1
threadpoolctl: 2.2.0
texttable: 1.6.4
testpath: 0.5.0
terminado: 0.12.1
tenacity: 8.0.1
tblib: 1.7.0
statsmodels: 0.12.2
sphinxcontrib-websupport: 1.2.4
sphinxcontrib-serializinghtml: 1.1.5
sphinxcontrib-qthelp: 1.0.3
sphinxcontrib-jsmath: 1.0.1
sphinxcontrib-htmlhelp: 2.0.0
sphinxcontrib-devhelp: 1.0.2
sphinxcontrib-applehelp: 1.0.2
sphinx: 3.5.4
sphinx-rtd-theme: 0.4.3
sphinx-autoapi: 1.8.3
soupsieve: 2.2.1
sortedcontainers: 2.4.0
snowballstemmer: 2.1.0
slicer: 0.0.7
sktime: 0.7.0
six: 1.16.0
shap: 0.39.0
setuptools: 58.0.4
send2trash: 1.8.0
```

(continues on next page)

(continued from previous page)

```
seaborn: 0.11.2
scipy: 1.7.1
scikit-optimize: 0.8.1
scikit-learn: 0.24.2
requirements-parser: 0.2.0
requests: 2.26.0
regex: 2021.8.28
recommonmark: 0.5.0
readthedocs-sphinx-ext: 2.1.4
pyzmq: 22.2.1
pyyaml: 5.4.1
pytz: 2021.1
python-dateutil: 2.8.2
pysistent: 0.18.0
pyparsing: 2.4.7
pygments: 2.10.0
pydata-sphinx-theme: 0.6.3
pyparser: 2.20
pyaml: 21.8.3
ptyprocess: 0.7.0
psutil: 5.8.0
prompt-toolkit: 3.0.20
prometheus-client: 0.11.0
pmdarima: 1.8.0
plotly: 5.3.1
pip: 21.2.4
pillow: 8.3.2
pickleshare: 0.7.5
pexpect: 4.8.0
patsy: 0.5.1
partd: 1.2.0
parso: 0.8.2
pandocfilters: 1.4.3
pandas: 1.3.2
packaging: 21.0
numpy: 1.21.2
numba: 0.53.0
notebook: 6.4.3
nltk: 3.6.2
nlp-primitives: 1.2.0
networkx: 2.5.1
nest-asyncio: 1.5.1
nbsphinx: 0.8.7
nbformat: 5.1.3
nbconvert: 6.1.0
nbclient: 0.5.4
msgpack: 1.0.2
mock: 1.0.1
mistune: 0.8.4
matplotlib: 3.4.3
matplotlib-inline: 0.1.3
markupsafe: 2.0.1
locket: 0.2.1
llvmlite: 0.36.0
lightgbm: 3.2.1
lazy-object-proxy: 1.6.0
kiwisolver: 1.3.2
```

(continues on next page)

(continued from previous page)

```
kaleido: 0.2.1
jupyterlab-widgets: 1.0.1
jupyterlab-pygments: 0.1.2
jupyter-core: 4.7.1
jupyter-client: 7.0.2
jsonschema: 3.2.0
joblib: 1.0.1
jinja2: 3.0.1
jedi: 0.18.0
ipywidgets: 7.6.4
ipython: 7.27.0
ipython-genutils: 0.2.0
ipykernel: 6.4.1
imbalanced-learn: 0.8.0
imagesize: 1.2.0
idna: 3.2
heapdict: 1.0.1
graphviz: 0.17
future: 0.18.2
fsspec: 2021.8.1
featuretools: 0.27.1
evalml: 0.32.1
entrypoints: 0.3
docutils: 0.16
distributed: 2021.9.0
defusedxml: 0.7.1
decorator: 4.4.2
debugpy: 1.4.3
dask: 2021.9.0
cython: 0.29.17
cyclcr: 0.10.0
commonmark: 0.8.1
colorama: 0.4.4
cloudpickle: 2.0.0
click: 8.0.1
charset-normalizer: 2.0.4
cffi: 1.14.6
certifi: 2021.5.30
category-encoders: 2.2.2
catboost: 0.26.1
bleach: 4.1.0
beautifulsoup4: 4.10.0
backcall: 0.2.0
babel: 2.9.1
attrs: 21.2.0
astroid: 2.6.6
argon2-cffi: 21.1.0
alabaster: 0.7.12
```

4.10 AutoMLSearch for time series problems

In this guide, we'll show how you can use EvalML to perform an automated search of machine learning pipelines for time series problems. Time series support is still being actively developed in EvalML so expect this page to improve over time.

4.10.1 But first, what is a time series?

A time series is a series of measurements taken at different moments in time ([Wikipedia](#)). The main difference between a time series dataset and a normal dataset is that the rows of a time series dataset are ordered with time. This relationship between the rows does not exist in non-time series datasets. In a non-time-series dataset, you can shuffle the rows and the dataset still has the same meaning. If you shuffle the rows of a time series dataset, the relationship between the rows is completely different!

4.10.2 What does AutoMLSearch for time series do?

In a machine learning setting, we are usually interested in using past values of the time series to predict future values. That is what EvalML's time series functionality is built to do.

4.10.3 Loading the data

In this guide, we work with daily minimum temperature recordings from Melbourne, Australia from the beginning of 1981 to end of 1990.

We start by loading the temperature data into two splits. The first split will be a training split consisting of data from 1981 to end of 1989. This is the data we'll use to find the best pipeline with AutoML. The second split will be a testing split consisting of data from 1990. This is the split we'll use to evaluate how well our pipeline generalizes on unseen data.

```
[1]: from evalml.demos import load_weather
X, y = load_weather()

[2]: train_dates, test_dates = X.Date < "1990-01-01", X.Date >= "1990-01-01"
X_train, y_train = X.ww.loc[train_dates], y.ww.loc[train_dates]
X_test, y_test = X.ww.loc[test_dates], y.ww.loc[test_dates]
```

Visualizing the training set

```
[3]: import plotly.graph_objects as go

[4]: data = [
    go.Scatter(
        x=X_train["Date"],
        y=y_train,
        mode="lines+markers",
        name="Temperature (C)",
        line=dict(color="#1f77b4"),
    )
]
# Let plotly pick the best date format.
```

(continues on next page)

(continued from previous page)

```

layout = go.Layout(
    title={"text": "Min Daily Temperature, Melbourne 1980-1989"},
    xaxis={"title": "Time"},
    yaxis={"title": "Temperature (C)"},
)

go.Figure(data=data, layout=layout)

```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

4.10.4 Running AutoMLSearch

AutoMLSearch for time series problems works very similarly to the other problem types with the exception that users need to pass in a new parameter called `problem_configuration`.

The `problem_configuration` is a dictionary specifying the following values:

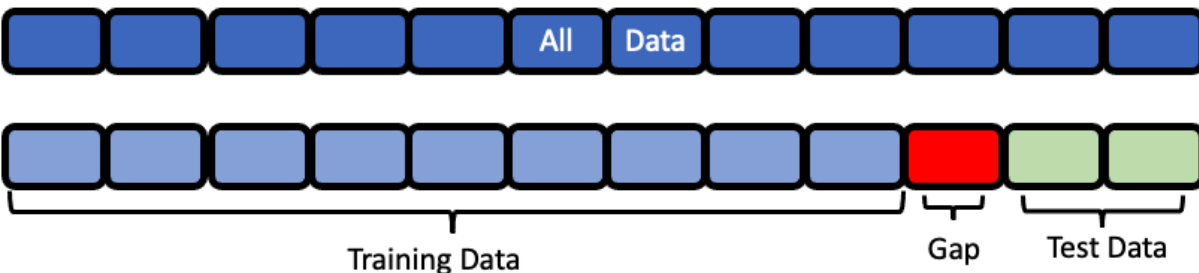
- **forecast_horizon:** The number of time periods we are trying to forecast. In this example, we're interested in predicting weather for the next 7 days, so the value is 7.
- **gap:** The number of time periods between the end of the training set and the start of the test set. For example, in our case we are interested in predicting the weather for the next 7 days with the data as it is "today", so the gap is 0. However, if we had to predict the weather for next Monday-Sunday with the data as it was on the previous Friday, the gap would be 2 (Saturday and Sunday separate Monday from Friday). It is important to select a value that matches the realistic delay between the forecast date and the most recently available data that can be used to make that forecast.
- **max_delay:** The maximum number of rows to look in the past from the current row in order to compute features. In our example, we'll say we can use the previous week's weather to predict the current week's.
- **date_index:** The column of the training dataset that contains the date corresponding to each observation. Currently, this parameter is only used by some time-series specific models so in this example, we are passing in `None`.

Note that the values of these parameters must be in the same units as the training/testing data.

Visualization of forecast horizon and gap

Forecast Horizon: 2

Gap: 1



```
[5]: from evalml.automl import AutoMLSearch

automl = AutoMLSearch(X_train, y_train, problem_type="time series regression",
                      max_batches=1,
                      problem_configuration={"gap": 0, "max_delay": 7,
                                           "forecast_horizon": 7, "date_index":
↳None},
                      allowed_model_families=["xgboost", "random_forest", "linear_
↳model", "extra_trees",
                                           "decision_tree"]
                      )

/home/docs/checkouts/readthedocs.org/user_builds/feature-labs-inc-evalml/envs/v0.32.1/
↳lib/python3.8/site-packages/evalml/automl/automl_search.py:464: UserWarning:

Time series support in evalml is still in beta, which means we are still actively_
↳building its core features. Please be mindful of that when running search().
```

```
[6]: automl.search()

/home/docs/checkouts/readthedocs.org/user_builds/feature-labs-inc-evalml/envs/v0.32.1/
↳lib/python3.8/site-packages/sklearn/linear_model/_coordinate_descent.py:530:
↳ConvergenceWarning:

Objective did not converge. You might want to increase the number of iterations.
↳Duality gap: 4089.24423870312, tolerance: 1.7012860284653468

/home/docs/checkouts/readthedocs.org/user_builds/feature-labs-inc-evalml/envs/v0.32.1/
↳lib/python3.8/site-packages/sklearn/linear_model/_coordinate_descent.py:530:
↳ConvergenceWarning:

Objective did not converge. You might want to increase the number of iterations.
↳Duality gap: 7638.417351298465, tolerance: 2.9630934475138115

/home/docs/checkouts/readthedocs.org/user_builds/feature-labs-inc-evalml/envs/v0.32.1/
↳lib/python3.8/site-packages/sklearn/linear_model/_coordinate_descent.py:530:
↳ConvergenceWarning:

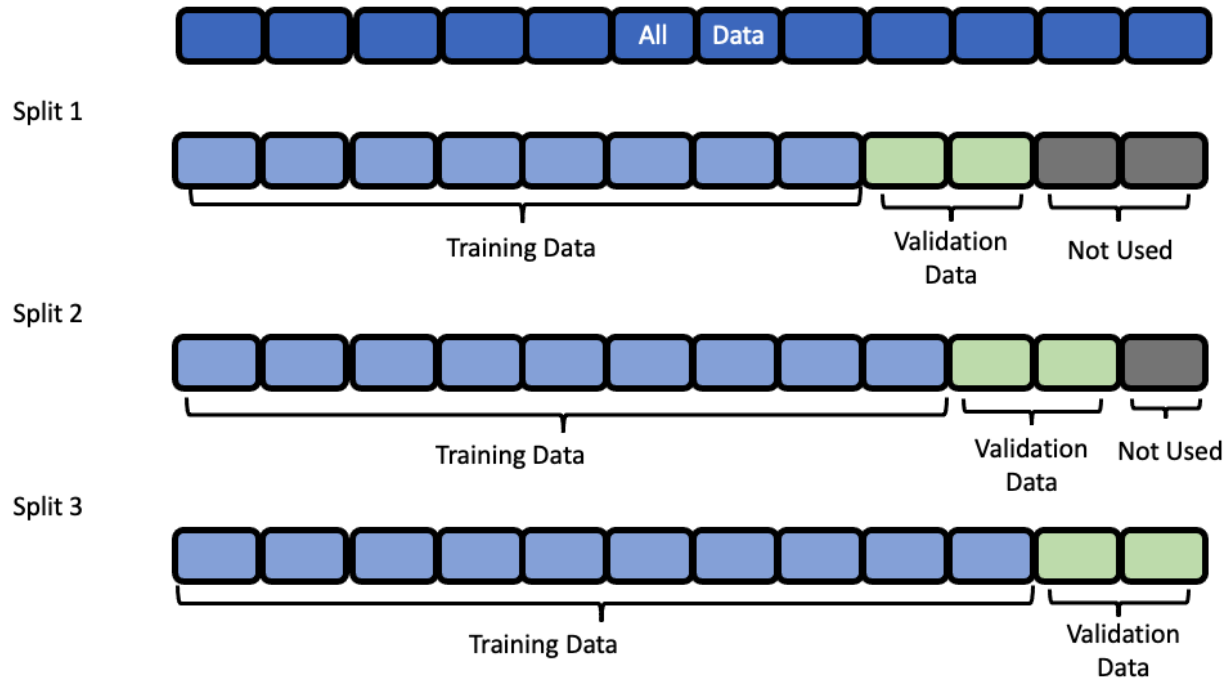
Objective did not converge. You might want to increase the number of iterations.
↳Duality gap: 11061.119581322855, tolerance: 4.09238672
```

4.10.5 Understanding what happened under the hood

This is great, AutoMLSearch is able to find a pipeline that scores an R2 value of 0.44 compared to a baseline pipeline that is only able to score 0.07. But how did it do that?

Data Splitting

EvalML uses [rolling origin cross validation](#) for time series problems. Basically, we take successive cuts of the training data while keeping the validation set size fixed. Note that the splits are not separated by `gap` number of units. This is because we need access to all the data to generate features for every row of the validation set. However, the feature engineering done by our pipelines respects the `gap` value. This is explained more in the [feature engineering section](#).



Baseline Pipeline

The most naive thing we can do in a time series problem is use the most recently available observation to predict the next observation. In our example, this means we'll use the measurement from 7 days ago as the prediction for the current date.

```
[7]: import pandas as pd
baseline = automl.get_pipeline(0)
baseline.fit(X_train, y_train)
naive_baseline_preds = baseline.predict_in_sample(X_test, y_test, objective=None,
                                                  X_train=X_train, y_train=y_train)
expected_preds = pd.concat([y_train.iloc[-7:], y_test]).shift(7).iloc[7:]
pd.testing.assert_series_equal(expected_preds, naive_baseline_preds)
```

Feature Engineering

EvalML uses the values of `gap`, `forecast_horizon`, and `max_delay` to calculate a “window” of allowed dates that can be used for engineering the features of each row in the validation/test set. The formula for computing the bounds of the window is:

$[t - (\text{max_delay} + \text{forecast_horizon} + \text{gap}), t - (\text{forecast_horizon} + \text{gap})]$

As an example, this is what the features for the first five days of August would look like in our current problem:

Forecast Horizon: 7

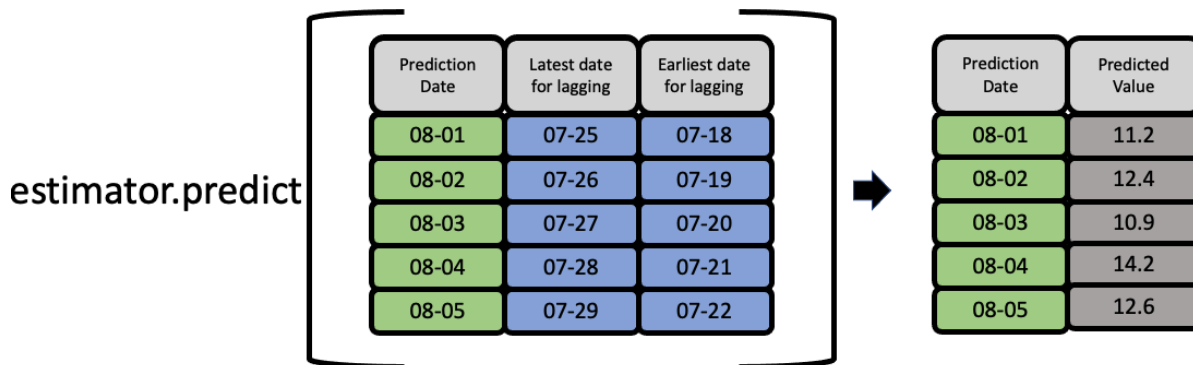
Gap: 0

Max Delay: 7

	Prediction Date	Latest date for lagging	Earliest date for lagging
■ Prediction point	08-01	07-25	07-18
■ Data before the forecast point	08-02	07-26	07-19
	08-03	07-27	07-20
	08-04	07-28	07-21
	08-05	07-29	07-22

The estimator then takes these features to generate predictions:

How the estimator generates predictions



4.10.6 Evaluate best pipeline on test data

Now that we have covered the mechanics of how EvalML runs `AutoMLSearch` for time series pipelines, we can compare the performance on the test set of the best pipeline found during search and the baseline pipeline.

```
[8]: pl = automl.best_pipeline
pl.fit(X_train, y_train)
best_pipeline_score = pl.score(X_test, y_test, ['R2'], X_train, y_train)['R2']
```

```
[9]: best_pipeline_score
```

```
[9]: 0.5502791890460579
```

```
[10]: baseline = automl.get_pipeline(0)
baseline.fit(X_train, y_train)
naive_baseline_score = baseline.score(X_test, y_test, ['R2'], X_train, y_train)['R2']
```

```
[11]: naive_baseline_score
```

```
[11]: 0.14924457197371788
```

The pipeline found by AutoMLSearch has a 268% improvement over the naive forecast!

```
[12]: automl.objective.calculate_percent_difference(best_pipeline_score, naive_baseline_
↪score)
```

```
[12]: 268.7096835541615
```

4.10.7 Visualize the predictions over time

```
[13]: from evalml.model_understanding import graph_prediction_vs_actual_over_time

fig = graph_prediction_vs_actual_over_time(pl, X_test, y_test, X_train, y_train,
↪dates=X_test['Date'])
fig
```

Data type cannot be displayed: application/vnd.plotly.v1+json, text/html

4.10.8 Predicting on unseen data

You'll notice that in the code snippets here, we use the `predict_in_sample` pipeline method as opposed to the usual `predict` method. What's the difference?

- `predict_in_sample` is used when the target value is known on the dates we are predicting on. This is true in cross validation. This method has an expected `y` parameter so that we can compute features using previous target values for all of the observations on the holdout set.
- `predict` is used when the target value is not known, e.g. the test dataset. The `y` parameter is not expected as only the target is observed in the training set. The test dataset must be separated by `gap` days from the training dataset. For the moment, the test set size must match `forecast_horizon`.

Here is an example of these two methods in action:

predict_in_sample

```
[14]: pl.predict_in_sample(X_test, y_test, objective=None, X_train=X_train, y_train=y_train)
[14]: 3285    14.222870
      3286    14.278846
      3287    14.022847
      3288    13.738781
      3289    13.530056
      ...
      3645    14.148007
      3646    13.927396
      3647    14.022316
      3648    14.045560
      3649    13.694097
      Name: Temp, Length: 365, dtype: float64
```

predict

```
[15]: pl.predict(X_test.iloc[:pl.forecast_horizon], objective=None, X_train=X_train, y_
      ↪train=y_train)
[15]: 3285    14.222870
      3286    14.278846
      3287    14.022847
      3288    13.738781
      3289    13.530056
      3290    13.988278
      3291    13.638807
      Name: Temp, dtype: float64
```

4.11 FAQ

4.11.1 Q: What is the difference between EvalML and other AutoML libraries?

EvalML optimizes machine learning pipelines on *custom practical objectives* instead of vague machine learning loss functions so that it will find the best pipelines for your specific needs. Furthermore, EvalML *pipelines* are able to take in all kinds of data (missing values, categorical, etc.) as long as the data are in a single table. EvalML also allows you to build your own pipelines with existing or custom components so you can have more control over the AutoML process. Moreover, EvalML also provides you with support in the form of *data checks* to ensure that you are aware of potential issues your data may cause with machine learning algorithms.

4.11.2 Q: How does EvalML handle missing values?

EvalML contains imputation components in its pipelines so that missing values are taken care of. EvalML optimizes over different types of imputation to search for the best possible pipeline. You can find more information about components [here](#) and in the API reference [here] (`../generated/evalml.pipelines.components.Imputer.ipynb`).

4.11.3 Q: How does EvalML handle categorical encoding?

EvalML provides a [one-hot-encoding component] ([../generated/evalml.pipelines.components.OneHotEncoder.ipynb](#)) in its pipelines for categorical variables. EvalML plans to support other encoders in the future.

4.11.4 Q: How does EvalML handle feature selection?

EvalML currently utilizes scikit-learn's [SelectFromModel](#) with a Random Forest classifier/regressor to handle feature selection. EvalML plans on supporting more feature selectors in the future. You can find more information in the API reference [\[here\]](#) ([../generated/evalml.pipelines.components.RFClassifierSelectFromModel.ipynb](#)).

4.11.5 Q: How is feature importance calculated?

Feature importance depends on the estimator used. Variable coefficients are used for regression-based estimators (Logistic Regression and Linear Regression) and Gini importance is used for tree-based estimators (Random Forest and XGBoost).

4.11.6 Q: How does hyperparameter tuning work?

EvalML tunes hyperparameters for its pipelines through Bayesian optimization. In the future we plan to support more optimization techniques such as random search.

4.11.7 Q: Can I create my own objective metric?

Yes you can! You can *create your own custom objective* so that EvalML optimizes the best model for your needs.

4.11.8 Q: How does EvalML avoid overfitting?

EvalML provides *data checks* to combat overfitting. Such data checks include detecting label leakage, unstable pipelines, hold-out datasets and cross validation. EvalML defaults to using Stratified K-Fold cross-validation for classification problems and K-Fold cross-validation for regression problems but allows you to utilize your own cross-validation methods as well.

4.11.9 Q: Can I create my own pipeline for EvalML?

Yes! EvalML allows you to create *custom pipelines* using modular components. This allows you to customize EvalML pipelines for your own needs or for AutoML.

4.11.10 Q: Does EvalML work with X algorithm?

EvalML is constantly improving and adding new components and will allow your own algorithms to be used as components in our pipelines.

API REFERENCE**5.1 Demo Datasets**

<i>load_breast_cancer</i>	Load breast cancer dataset. Binary classification problem.
<i>load_diabetes</i>	Load diabetes dataset. Regression problem
<i>load_fraud</i>	Load credit card fraud dataset.
<i>load_wine</i>	Load wine dataset. Multiclass problem.
<i>load_churn</i>	Load credit card fraud dataset.

5.2 Preprocessing

Utilities to preprocess data before using evalml.

<i>load_data</i>	Load features and target from file.
<i>target_distribution</i>	Get the target distributions.
<i>number_of_features</i>	Get the number of features of each specific dtype in a DataFrame.
<i>split_data</i>	Splits data into train and test sets.

5.3 Exceptions

<i>MethodPropertyNotFoundError</i>	Exception to raise when a class does not have an expected method or property.
<i>PipelineNotFoundError</i>	An exception raised when a particular pipeline is not found in automl search results
<i>ObjectiveNotFoundError</i>	Exception to raise when specified objective does not exist.
<i>MissingComponentError</i>	An exception raised when a component is not found in <code>all_components()</code>
<i>ComponentNotYetFittedError</i>	An exception to be raised when <code>predict/predict_proba/transform</code> is called on a component without fitting first.

continues on next page

Table 3 – continued from previous page

<i>PipelineNotYetFittedError</i>	An exception to be raised when predict/predict_proba/transform is called on a pipeline without fitting first.
<i>AutoMLSearchException</i>	Exception raised when all pipelines in an automl batch return a score of NaN for the primary objective.
<i>EnsembleMissingPipelinesError</i>	An exception raised when an ensemble is missing <i>estimators</i> (list) as a parameter.
<i>PipelineScoreError</i>	An exception raised when a pipeline errors while scoring any objective in a list of objectives.
<i>DataCheckInitError</i>	Exception raised when a data check can't initialize with the parameters given.
<i>NullsInColumnWarning</i>	Warning thrown when there are null values in the column of interest

5.4 AutoML

5.4.1 AutoML Search Interface

<i>AutoMLSearch</i>	Automated Pipeline search.
---------------------	----------------------------

5.4.2 AutoML Utils

<i>search</i>	Given data and configuration, run an automl search.
<i>get_default_primary_search_objective</i>	Get the default primary search objective for a problem type.
<i>make_data_splitter</i>	Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.

5.4.3 AutoML Algorithm Classes

<i>AutoMLAlgorithm</i>	Base class for the AutoML algorithms which power EvalML.
<i>IterativeAlgorithm</i>	An automl algorithm which first fits a base round of pipelines with default parameters, then does a round of parameter tuning on each pipeline in order of performance.

5.4.4 AutoML Callbacks

<code>silent_error_callback</code>	No-op.
<code>log_error_callback</code>	Logs the exception thrown as an error. Will not throw. This is the default behavior for AutoMLSearch.
<code>raise_error_callback</code>	Raises the exception thrown by the AutoMLSearch object. Also logs the exception as an error.

5.4.5 AutoML Engines

<code>SequentialEngine</code>	The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.
<code>CFEngine</code>	The concurrent.futures (CF) engine
<code>DaskEngine</code>	The dask engine

5.5 Pipelines

5.5.1 Pipeline Base Classes

<code>PipelineBase</code>	Machine learning pipeline made out of transformers and an Estimator.
<code>ClassificationPipeline</code>	Pipeline subclass for all classification pipelines.
<code>BinaryClassificationPipeline</code>	Pipeline subclass for all binary classification pipelines.
<code>MulticlassClassificationPipeline</code>	Pipeline subclass for all multiclass classification pipelines.
<code>RegressionPipeline</code>	Pipeline subclass for all regression pipelines.
<code>TimeSeriesClassificationPipeline</code>	Pipeline base class for time series classification problems.
<code>TimeSeriesBinaryClassificationPipeline</code>	Pipeline base class for time series binary classification problems.
<code>TimeSeriesMulticlassClassificationPipeline</code>	Pipeline base class for time series multiclass classification problems.
<code>TimeSeriesRegressionPipeline</code>	Pipeline base class for time series regression problems.

5.5.2 Pipeline Utils

<code>make_pipeline</code>	Given input data, target data, an estimator class and the problem type,
<code>generate_pipeline_code</code>	Creates and returns a string that contains the Python imports and code required for running the EvalML pipeline.

5.6 Component Graphs

<i>ComponentGraph</i>	Component graph for a pipeline as a directed acyclic graph (DAG).
-----------------------	---

5.7 Components

5.7.1 Component Base Classes

Components represent a step in a pipeline.

<i>ComponentBase</i>	Base class for all components.
<i>Transformer</i>	A component that may or may not need fitting that transforms data.
<i>Estimator</i>	A component that fits and predicts given data.

5.7.2 Component Utils

<i>allowed_model_families</i>	List the model types allowed for a particular problem type.
<i>get_estimators</i>	Returns the estimators allowed for a particular problem type.
<i>generate_component_code</i>	Creates and returns a string that contains the Python imports and code required for running the EvalML component.

5.7.3 Transformers

Transformers are components that take in data as input and output transformed data.

<i>DropColumns</i>	Drops specified columns in input data.
<i>SelectColumns</i>	Selects specified columns in input data.
<i>SelectByType</i>	Selects columns by specified Woodwork logical type or semantic tag in input data.
<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.
<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
<i>Imputer</i>	Imputes missing data according to a specified imputation strategy.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.

continues on next page

Table 14 – continued from previous page

<i>StandardScaler</i>	A transformer that standardizes input features by removing the mean and scaling to unit variance.
<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.
<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
<i>DropNullColumns</i>	Transformer to drop features whose percentage of NaN values exceeds a specified threshold.
<i>DateTimeFeaturizer</i>	Transformer that can automatically extract features from datetime columns.
<i>TextFeaturizer</i>	Transformer that can automatically featurize text columns using featuretools' nlp_primitives.
<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
<i>PolynomialDetrender</i>	Removes trends from time series by fitting a polynomial to the data.
<i>Undersampler</i>	Initializes an undersampling transformer to downsample the majority classes in the dataset.
<i>Oversampler</i>	SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMO-TENC based on inputs to the component.

5.7.4 Estimators

Classifiers

Classifiers are components that output a predicted class label.

<i>CatBoostClassifier</i>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.
<i>ElasticNetClassifier</i>	Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.
<i>ExtraTreesClassifier</i>	Extra Trees Classifier.
<i>RandomForestClassifier</i>	Random Forest Classifier.
<i>LightGBMClassifier</i>	LightGBM Classifier.
<i>LogisticRegressionClassifier</i>	Logistic Regression Classifier.
<i>XGBoostClassifier</i>	XGBoost Classifier.
<i>BaselineClassifier</i>	Classifier that predicts using the specified strategy.
<i>SklearnStackedEnsembleClassifier</i>	Scikit-learn Stacked Ensemble Classifier.
<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
<i>KNeighborsClassifier</i>	K-Nearest Neighbors Classifier.
<i>SVMClassifier</i>	Support Vector Machine Classifier.

Regressors

Regressors are components that output a predicted target value.

<i>ARIMAREgressor</i>	Autoregressive Integrated Moving Average Model.
<i>CatBoostRegressor</i>	CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.
<i>ElasticNetRegressor</i>	Elastic Net Regressor.
<i>LinearRegressor</i>	Linear Regressor.
<i>ExtraTreesRegressor</i>	Extra Trees Regressor.
<i>RandomForestRegressor</i>	Random Forest Regressor.
<i>XGBoostRegressor</i>	XGBoost Regressor.
<i>BaselineRegressor</i>	Baseline regressor that uses a simple strategy to make predictions.
<i>TimeSeriesBaselineEstimator</i>	Time series estimator that predicts using the naive forecasting approach.
<i>SklearnStackedEnsembleRegressor</i>	Scikit-learn Stacked Ensemble Regressor.
<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>SVMRegressor</i>	Support Vector Machine Regressor.

5.8 Model Understanding

5.8.1 Utility Methods

<i>confusion_matrix</i>	Confusion matrix for binary and multiclass classification.
<i>normalize_confusion_matrix</i>	Normalizes a confusion matrix.
<i>precision_recall_curve</i>	Given labels and binary classifier predicted probabilities, compute and return the data representing a precision-recall curve.
<i>roc_curve</i>	Given labels and classifier predicted probabilities, compute and return the data representing a Receiver Operating Characteristic (ROC) curve. Works with binary or multiclass problems.
<i>calculate_permutation_importance</i>	Calculates permutation importance for features.
<i>calculate_permutation_importance_one_column</i>	Calculates permutation importance for one column in the original dataframe.
<i>binary_objective_vs_threshold</i>	Computes objective score as a function of potential binary classification
<i>get_prediction_vs_actual_over_time_data</i>	Get the data needed for the prediction_vs_actual_over_time plot.
<i>partial_dependence</i>	Calculates one or two-way partial dependence. If a single integer or
<i>get_prediction_vs_actual_data</i>	Combines y_true and y_pred into a single dataframe and adds a column for outliers. Used in <i>graph_prediction_vs_actual()</i> .
<i>get_linear_coefficients</i>	Returns a dataframe showing the features with the greatest predictive power for a linear model.

continues on next page

Table 17 – continued from previous page

<i>t_sne</i>	Get the transformed output after fitting X to the embedded space using t-SNE.
--------------	---

5.8.2 Graph Utility Methods

<i>graph_precision_recall_curve</i>	Generate and display a precision-recall plot.
<i>graph_roc_curve</i>	Generate and display a Receiver Operating Characteristic (ROC) plot for binary and multiclass classification problems.
<i>graph_confusion_matrix</i>	Generate and display a confusion matrix plot.
<i>graph_permutation_importance</i>	Generate a bar graph of the pipeline’s permutation importance.
<i>graph_binary_objective_vs_threshold</i>	Generates a plot graphing objective score vs. decision thresholds for a fitted binary classification pipeline.
<i>graph_prediction_vs_actual</i>	Generate a scatter plot comparing the true and predicted values. Used for regression plotting
<i>graph_prediction_vs_actual_over_time</i>	Plot the target values and predictions against time on the x-axis.
<i>graph_partial_dependence</i>	Create an one-way or two-way partial dependence plot. Passing a single integer or
<i>graph_t_sne</i>	Plot high dimensional data into lower dimensional space using t-SNE .

5.8.3 Prediction Explanations

<i>explain_predictions</i>	Creates a report summarizing the top contributing features for each data point in the input features.
<i>explain_predictions_best_worst</i>	Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

5.9 Objectives

5.9.1 Objective Base Classes

<i>ObjectiveBase</i>	Base class for all objectives.
<i>BinaryClassificationObjective</i>	Base class for all binary classification objectives.
<i>MulticlassClassificationObjective</i>	Base class for all multiclass classification objectives.
<i>RegressionObjective</i>	Base class for all regression objectives.

5.9.2 Domain-Specific Objectives

<i>FraudCost</i>	Score the percentage of money lost of the total transaction amount process due to fraud.
<i>LeadScoring</i>	Lead scoring.
<i>CostBenefitMatrix</i>	Score using a cost-benefit matrix. Scores quantify the benefits of a given value, so greater numeric

5.9.3 Classification Objectives

<i>AccuracyBinary</i>	Accuracy score for binary classification.
<i>AccuracyMulticlass</i>	Accuracy score for multiclass classification.
<i>AUC</i>	AUC score for binary classification.
<i>AUCMacro</i>	AUC score for multiclass classification using macro averaging.
<i>AUCMicro</i>	AUC score for multiclass classification using micro averaging.
<i>AUCWeighted</i>	AUC Score for multiclass classification using weighted averaging.
<i>Gini</i>	Gini coefficient for binary classification.
<i>BalancedAccuracyBinary</i>	Balanced accuracy score for binary classification.
<i>BalancedAccuracyMulticlass</i>	Balanced accuracy score for multiclass classification.
<i>F1</i>	F1 score for binary classification.
<i>F1Micro</i>	F1 score for multiclass classification using micro averaging.
<i>F1Macro</i>	F1 score for multiclass classification using macro averaging.
<i>F1Weighted</i>	F1 score for multiclass classification using weighted averaging.
<i>LogLossBinary</i>	Log Loss for binary classification.
<i>LogLossMulticlass</i>	Log Loss for multiclass classification.
<i>MCCBinary</i>	Matthews correlation coefficient for binary classification.
<i>MCCMulticlass</i>	Matthews correlation coefficient for multiclass classification.
<i>Precision</i>	Precision score for binary classification.
<i>PrecisionMicro</i>	Precision score for multiclass classification using micro averaging.
<i>PrecisionMacro</i>	Precision score for multiclass classification using macro averaging.
<i>PrecisionWeighted</i>	Precision score for multiclass classification using weighted averaging.
<i>Recall</i>	Recall score for binary classification.
<i>RecallMicro</i>	Recall score for multiclass classification using micro averaging.
<i>RecallMacro</i>	Recall score for multiclass classification using macro averaging.
<i>RecallWeighted</i>	Recall score for multiclass classification using weighted averaging.

5.9.4 Regression Objectives

<i>R2</i>	Coefficient of determination for regression.
<i>MAE</i>	Mean absolute error for regression.
<i>MAPE</i>	Mean absolute percentage error for time series regression. Scaled by 100 to return a percentage.
<i>MSE</i>	Mean squared error for regression.
<i>MeanSquaredLogError</i>	Mean squared log error for regression.
<i>MedianAE</i>	Median absolute error for regression.
<i>MaxError</i>	Maximum residual error for regression.
<i>ExpVariance</i>	Explained variance score for regression.
<i>RootMeanSquaredError</i>	Root mean squared error for regression.
<i>RootMeanSquaredLogError</i>	Root mean squared log error for regression.

5.9.5 Objective Utils

<i>get_all_objective_names</i>	Get a list of the names of all objectives.
<i>get_core_objectives</i>	Returns all core objective instances associated with the given problem type.
<i>get_core_objective_names</i>	Get a list of all valid core objectives.
<i>get_non_core_objectives</i>	Get non-core objective classes.
<i>get_objective</i>	Returns the Objective class corresponding to a given objective name.

5.10 Problem Types

<i>handle_problem_types</i>	Handles <i>problem_type</i> by either returning the ProblemTypes or converting from a str.
<i>detect_problem_type</i>	Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)
<i>ProblemTypes</i>	Enum defining the supported types of machine learning problems.

5.11 Model Family

<i>handle_model_family</i>	Handles <i>model_family</i> by either returning the ModelFamily or converting from a string
<i>ModelFamily</i>	Enum for family of machine learning models.

5.12 Tuners

<i>Tuner</i>	Defines API for base Tuner classes.
<i>SKOptTuner</i>	Bayesian Optimizer.
<i>GridSearchTuner</i>	Grid Search Optimizer, which generates all of the possible points to search for using a grid.
<i>RandomSearchTuner</i>	Random Search Optimizer.

5.13 Data Checks

5.13.1 Data Check Classes

<i>DataCheck</i>	Base class for all data checks. Data checks are a set of heuristics used to determine if there are problems with input data.
<i>InvalidTargetDataCheck</i>	Checks if the target data contains missing or invalid values.
<i>HighlyNullDataCheck</i>	Checks if there are any highly-null columns and rows in the input.
<i>IDColumnsDataCheck</i>	Check if any of the features are likely to be ID columns.
<i>TargetLeakageDataCheck</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.
<i>OutliersDataCheck</i>	Checks if there are any outliers in input data by using IQR to determine score anomalies. Columns with score anomalies are considered to contain outliers.
<i>NoVarianceDataCheck</i>	Check if the target or any of the features have no variance.
<i>ClassImbalanceDataCheck</i>	Check if any of the target labels are imbalanced, or if the number of values for each target are below 2 times the number of CV folds. Use for classification problems.
<i>MulticollinearityDataCheck</i>	Check if any set features are likely to be multicollinear.
<i>DateTimeNaNDataCheck</i>	Checks each column in the input for datetime features and will issue an error if NaN values are present.
<i>NaturalLanguageNaNDataCheck</i>	Checks each column in the input for natural language features and will issue an error if NaN values are present.
<i>DateTimeFormatDataCheck</i>	Checks if the datetime column has equally spaced intervals and is monotonically increasing or decreasing in order
<i>DataChecks</i>	A collection of data checks.
<i>DefaultDataChecks</i>	A collection of basic data checks that is used by AutoML by default.

5.13.2 Data Check Messages

<i>DataCheckMessage</i>	Base class for a message returned by a DataCheck, tagged by name.
<i>DataCheckError</i>	DataCheckMessage subclass for errors returned by data checks.
<i>DataCheckWarning</i>	DataCheckMessage subclass for warnings returned by data checks.

5.13.3 Data Check Message Types

<i>DataCheckMessageType</i>	Enum for type of data check message: WARNING or ERROR.
-----------------------------	--

5.13.4 Data Check Message Codes

<i>DataCheckMessageCode</i>	Enum for data check message code.
-----------------------------	-----------------------------------

5.14 Utils

5.14.1 General Utils

<i>import_or_raise</i>	Attempts to import the requested library by name.
<i>convert_to_seconds</i>	Converts a string describing a length of time to its length in seconds.
<i>get_random_state</i>	Generates a <code>numpy.random.RandomState</code> instance using seed.
<i>get_random_seed</i>	Given a <code>numpy.random.RandomState</code> object, generate an int representing a seed value for another random number generator. Or, if given an int, return that int.
<i>pad_with_nans</i>	Pad the beginning <code>num_to_pad</code> rows with nans.
<i>drop_rows_with_nans</i>	Drop rows that have any NaNs in all dataframes or series.
<i>infer_feature_types</i>	Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.
<i>save_plot</i>	Saves fig to filepath if specified, or to a default location if not.
<i>is_all_numeric</i>	Checks if the given DataFrame contains only numeric values
<i>get_importable_subclasses</i>	Get importable subclasses of a base class. Used to list all of our

Evalml

Subpackages

Automl

Subpackages

automl_algorithm

Submodules

automl_algorithm

Module Contents

Classes Summary

<i>AutoMLAlgorithm</i>	Base class for the AutoML algorithms which power EvalML.
------------------------	--

Exceptions Summary

Contents

```
class evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm(allowed_pipelines=None,
                                                                    cus-
                                                                    tom_hyperparameters=None,
                                                                    max_iterations=None,
                                                                    tuner_class=None,
                                                                    ran-
                                                                    dom_seed=0)
```

Base class for the AutoML algorithms which power EvalML.

This class represents an automated machine learning (AutoML) algorithm. It encapsulates the decision-making logic behind an automl search, by both deciding which pipelines to evaluate next and by deciding what set of parameters to configure the pipeline with.

To use this interface, you must define a `next_batch` method which returns the next group of pipelines to evaluate on the training data. That method may access state and results recorded from the previous batches, although that information is not tracked in a general way in this base class. Overriding `add_result` is a convenient way to record pipeline evaluation info if necessary.

Parameters

- **allowed_pipelines** (*list(class)*) – A list of PipelineBase subclasses indicating the pipelines allowed in the search. The default of None indicates all pipelines for this problem type are allowed.

- **custom_hyperparameters** (*dict*) – Custom hyperparameter ranges specified for pipelines to iterate over.
- **max_iterations** (*int*) – The maximum number of iterations to be evaluated.
- **tuner_class** (*class*) – A subclass of Tuner, to be used to find parameters for each pipeline. The default of None indicates the SKOptTuner will be used.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Methods

<code>add_result</code>	Register results from evaluating a pipeline
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

add_result (*self*, *score_to_minimize*, *pipeline*, *trained_pipeline_results*)

Register results from evaluating a pipeline

Parameters

- **score_to_minimize** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **pipeline** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **trained_pipeline_results** (*dict*) – Results from training a pipeline.

property batch_number (*self*)

Returns the number of batches which have been recommended so far.

abstract next_batch (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

property pipeline_number (*self*)

Returns the number of pipelines which have been recommended so far.

exception evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithmException

Exception raised when an error is encountered during the computation of the automl algorithm

default_algorithm

Module Contents

Classes Summary

<code>DefaultAlgorithm</code>	An automl algorithm that consists of two modes: fast and long, where fast is a subset of long.
-------------------------------	--

Contents

```
class evalml.automl.automl_algorithm.default_algorithm.DefaultAlgorithm(X,  
                                                                    y,  
                                                                    prob-  
                                                                    lem_type,  
                                                                    sam-  
                                                                    pler_name,  
                                                                    tuner_class=None,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    pipeline_params=None,  
                                                                    cus-  
                                                                    tom_hyperparameters=None,  
                                                                    n_jobs=-  
                                                                    1,  
                                                                    text_in_ensembling=None,  
                                                                    top_n=3,  
                                                                    num_long_explore_pipeline=  
                                                                    num_long_pipelines_per_b
```

An automl algorithm that consists of two modes: fast and long, where fast is a subset of long.

1. Naive pipelines:

- a. run baseline with default preprocessing pipeline
- b. run naive linear model with default preprocessing pipeline
- c. run basic RF pipeline with default preprocessing pipeline

2. Naive pipelines with feature selection

- a. subsequent pipelines will use the selected features with a SelectedColumns transformer

At this point we have a single pipeline candidate for preprocessing and feature selection

3. Pipelines with preprocessing components:

- a. scan rest of estimators (our current batch 1).

4. First ensembling run

Fast mode ends here. Begin long mode.

6. Run top 3 estimators:

- a. Generate 50 random parameter sets. Run all 150 in one batch

7. Second ensembling run

8. Repeat these indefinitely until stopping criterion is met:

- a. For each of the previous top 3 estimators, sample 10 parameters from the tuner. Run all 30 in one batch
- b. Run ensembling

Methods

<code>add_result</code>	Register results from evaluating a pipeline. In batch number 2, the selected column names from the feature selector are taken to be used in a column selector. Information regarding the best pipeline is updated here as well.
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

add_result (*self*, *score_to_minimize*, *pipeline*, *trained_pipeline_results*)

Register results from evaluating a pipeline. In batch number 2, the selected column names from the feature selector are taken to be used in a column selector. Information regarding the best pipeline is updated here as well.

Parameters

- **score_to_minimize** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **pipeline** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **trained_pipeline_results** (*dict*) – Results from training a pipeline.

property batch_number (*self*)

Returns the number of batches which have been recommended so far.

next_batch (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

property pipeline_number (*self*)

Returns the number of pipelines which have been recommended so far.

iterative_algorithm

Module Contents

Classes Summary

<code>IterativeAlgorithm</code>	An automl algorithm which first fits a base round of pipelines with default parameters, then does a round of parameter tuning on each pipeline in order of performance.
---------------------------------	---

Contents

```
class evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm(allowed_pipelines=None,
max_iterations=None,
tuner_class=None,
random_seed=0,
pipelines_per_batch=5,
n_jobs=None,

1,
number_features=None,
ensembling=False,
text_in_ensembling=False,
pipeline_params=None,
custom_hyperparameters=None,
_ensemble=None,
_tuning=None,
_maximization=None,
tor_family_order=None)
```

An automl algorithm which first fits a base round of pipelines with default parameters, then does a round of parameter tuning on each pipeline in order of performance.

Parameters

- **allowed_pipelines** (*list (class)*) – A list of PipelineBase instances indicating the pipelines allowed in the search. The default of None indicates all pipelines for this problem type are allowed.
- **max_iterations** (*int*) – The maximum number of iterations to be evaluated.
- **tuner_class** (*class*) – A subclass of Tuner, to be used to find parameters for each pipeline. The default of None indicates the SKOptTuner will be used.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **pipelines_per_batch** (*int*) – The number of pipelines to be evaluated in each batch, after the first batch. Defaults to 5.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to None.
- **number_features** (*int*) – The number of columns in the input features. Defaults to None.
- **ensembling** (*boolean*) – If True, runs ensembling in a separate batch after every allowed pipeline class has been iterated over. Defaults to False.
- **text_in_ensembling** (*boolean*) – If True and ensembling is True, then n_jobs will be set to 1 to avoid downstream sklearn stacking issues related to nltk. Defaults to None.
- **pipeline_params** (*dict or None*) – Pipeline-level parameters that should be passed to the proposed pipelines. Defaults to None.
- **custom_hyperparameters** (*dict or None*) – Custom hyperparameter ranges specified for pipelines to iterate over. Defaults to None.

- **`_estimator_family_order`** (*list(ModelFamily) or None*) – specify the sort order for the first batch. Defaults to None, which uses `_ESTIMATOR_FAMILY_ORDER`.

Methods

<code>add_result</code>	Register results from evaluating a pipeline
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

`add_result` (*self, score_to_minimize, pipeline, trained_pipeline_results*)

Register results from evaluating a pipeline

Parameters

- **`score_to_minimize`** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **`pipeline`** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **`trained_pipeline_results`** (*dict*) – Results from training a pipeline.

`property batch_number` (*self*)

Returns the number of batches which have been recommended so far.

`next_batch` (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

`property pipeline_number` (*self*)

Returns the number of pipelines which have been recommended so far.

Package Contents

Classes Summary

<code>AutoMLAlgorithm</code>	Base class for the AutoML algorithms which power EvalML.
<code>DefaultAlgorithm</code>	An automl algorithm that consists of two modes: fast and long, where fast is a subset of long.
<code>IterativeAlgorithm</code>	An automl algorithm which first fits a base round of pipelines with default parameters, then does a round of parameter tuning on each pipeline in order of performance.

Exceptions Summary

Contents

```
class evalml.automl.automl_algorithm.AutoMLAlgorithm(allowed_pipelines=None, custom_hyperparameters=None, max_iterations=None, tuner_class=None, random_seed=0)
```

Base class for the AutoML algorithms which power EvalML.

This class represents an automated machine learning (AutoML) algorithm. It encapsulates the decision-making logic behind an automl search, by both deciding which pipelines to evaluate next and by deciding what set of parameters to configure the pipeline with.

To use this interface, you must define a `next_batch` method which returns the next group of pipelines to evaluate on the training data. That method may access state and results recorded from the previous batches, although that information is not tracked in a general way in this base class. Overriding `add_result` is a convenient way to record pipeline evaluation info if necessary.

Parameters

- **allowed_pipelines** (*list(class)*) – A list of PipelineBase subclasses indicating the pipelines allowed in the search. The default of None indicates all pipelines for this problem type are allowed.
- **custom_hyperparameters** (*dict*) – Custom hyperparameter ranges specified for pipelines to iterate over.
- **max_iterations** (*int*) – The maximum number of iterations to be evaluated.
- **tuner_class** (*class*) – A subclass of Tuner, to be used to find parameters for each pipeline. The default of None indicates the SKOptTuner will be used.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Methods

<code>add_result</code>	Register results from evaluating a pipeline
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

add_result (*self, score_to_minimize, pipeline, trained_pipeline_results*)

Register results from evaluating a pipeline

Parameters

- **score_to_minimize** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **pipeline** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **trained_pipeline_results** (*dict*) – Results from training a pipeline.

property `batch_number` (*self*)

Returns the number of batches which have been recommended so far.

abstract `next_batch` (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

property `pipeline_number` (*self*)

Returns the number of pipelines which have been recommended so far.

exception `evalml.automl.automl_algorithm.AutoMLAlgorithmException`

Exception raised when an error is encountered during the computation of the automl algorithm

class `evalml.automl.automl_algorithm.DefaultAlgorithm` (*X*, *y*, *problem_type*,
sampler_name,
tuner_class=None,
random_seed=0,
pipeline_params=None, *custom_hyperparameters*=None,
n_jobs=-1,
text_in_ensembling=None,
top_n=3,
num_long_explore_pipelines=50,
num_long_pipelines_per_batch=10)

An automl algorithm that consists of two modes: fast and long, where fast is a subset of long.

1. Naive pipelines:

- a. run baseline with default preprocessing pipeline
- b. run naive linear model with default preprocessing pipeline
- c. run basic RF pipeline with default preprocessing pipeline

2. Naive pipelines with feature selection

- a. subsequent pipelines will use the selected features with a SelectedColumns transformer

At this point we have a single pipeline candidate for preprocessing and feature selection

3. Pipelines with preprocessing components:

- a. scan rest of estimators (our current batch 1).

4. First ensembling run

Fast mode ends here. Begin long mode.

6. Run top 3 estimators:

- a. Generate 50 random parameter sets. Run all 150 in one batch

7. Second ensembling run

8. Repeat these indefinitely until stopping criterion is met:

- a. For each of the previous top 3 estimators, sample 10 parameters from the tuner. Run all 30 in one batch
- b. Run ensembling

Methods

<code>add_result</code>	Register results from evaluating a pipeline. In batch number 2, the selected column names from the feature selector are taken to be used in a column selector. Information regarding the best pipeline is updated here as well.
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

add_result (*self*, *score_to_minimize*, *pipeline*, *trained_pipeline_results*)

Register results from evaluating a pipeline. In batch number 2, the selected column names from the feature selector are taken to be used in a column selector. Information regarding the best pipeline is updated here as well.

Parameters

- **score_to_minimize** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **pipeline** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **trained_pipeline_results** (*dict*) – Results from training a pipeline.

property batch_number (*self*)

Returns the number of batches which have been recommended so far.

next_batch (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

property pipeline_number (*self*)

Returns the number of pipelines which have been recommended so far.

```
class evalml.automl.automl_algorithm.IterativeAlgorithm (allowed_pipelines=None,  
                                                    max_iterations=None,  
                                                    tuner_class=None,  
                                                    random_seed=0,  
                                                    pipelines_per_batch=5,  
                                                    n_jobs=-1, number_features=None,  
                                                    ensembling=False,  
                                                    text_in_ensembling=False,  
                                                    pipeline_params=None,  
                                                    custom_hyperparameters=None,  
                                                    estimator_family_order=None)
```

An automl algorithm which first fits a base round of pipelines with default parameters, then does a round of parameter tuning on each pipeline in order of performance.

Parameters

- **allowed_pipelines** (*list(class)*) – A list of PipelineBase instances indicating the pipelines allowed in the search. The default of None indicates all pipelines for this problem type are allowed.
- **max_iterations** (*int*) – The maximum number of iterations to be evaluated.
- **tuner_class** (*class*) – A subclass of Tuner, to be used to find parameters for each pipeline. The default of None indicates the SKOptTuner will be used.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **pipelines_per_batch** (*int*) – The number of pipelines to be evaluated in each batch, after the first batch. Defaults to 5.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to None.
- **number_features** (*int*) – The number of columns in the input features. Defaults to None.
- **ensembling** (*boolean*) – If True, runs ensembling in a separate batch after every allowed pipeline class has been iterated over. Defaults to False.
- **text_in_ensembling** (*boolean*) – If True and ensembling is True, then n_jobs will be set to 1 to avoid downstream sklearn stacking issues related to nltk. Defaults to None.
- **pipeline_params** (*dict or None*) – Pipeline-level parameters that should be passed to the proposed pipelines. Defaults to None.
- **custom_hyperparameters** (*dict or None*) – Custom hyperparameter ranges specified for pipelines to iterate over. Defaults to None.
- **_estimator_family_order** (*list(ModelFamily) or None*) – specify the sort order for the first batch. Defaults to None, which uses _ESTIMATOR_FAMILY_ORDER.

Methods

<code>add_result</code>	Register results from evaluating a pipeline
<code>batch_number</code>	Returns the number of batches which have been recommended so far.
<code>next_batch</code>	Get the next batch of pipelines to evaluate
<code>pipeline_number</code>	Returns the number of pipelines which have been recommended so far.

add_result (*self, score_to_minimize, pipeline, trained_pipeline_results*)

Register results from evaluating a pipeline

Parameters

- **score_to_minimize** (*float*) – The score obtained by this pipeline on the primary objective, converted so that lower values indicate better pipelines.
- **pipeline** (*PipelineBase*) – The trained pipeline object which was used to compute the score.
- **trained_pipeline_results** (*dict*) – Results from training a pipeline.

property batch_number (*self*)

Returns the number of batches which have been recommended so far.

next_batch (*self*)

Get the next batch of pipelines to evaluate

Returns a list of instances of PipelineBase subclasses, ready to be trained and evaluated.

Return type list(PipelineBase)

property pipeline_number (*self*)

Returns the number of pipelines which have been recommended so far.

engine

Submodules

cf_engine

Module Contents

Classes Summary

<i>CFClient</i>	Custom CFClient API to match Dask's CFClient and allow context management.
<i>CFComputation</i>	A Future-like wrapper around jobs created by the CFEngine.
<i>CFEngine</i>	The concurrent.futures (CF) engine

Contents

class evalml.automl.engine.cf_engine.**CFClient** (*pool*)

Custom CFClient API to match Dask's CFClient and allow context management.

Methods

<i>close</i>	Closes the underlying Executor.
<i>is_closed</i>	
<i>submit</i>	Pass through to imitate Dask's Client API.

close (*self*)

Closes the underlying Executor.

property is_closed (*self*)

submit (*self*, *args, **kwargs)

Pass through to imitate Dask's Client API.

class evalml.automl.engine.cf_engine.**CFComputation** (*future*)

A Future-like wrapper around jobs created by the CFEngine.

Methods

<i>cancel</i>	Cancel the current computation.
---------------	---------------------------------

continues on next page

Table 47 – continued from previous page

<i>done</i>	returns Whether the computation is done.
<i>get_result</i>	Gets the computation result.
<i>is_cancelled</i>	returns Returns whether computation was cancelled.

cancel (*self*)

Cancel the current computation.

Returns

False if the call is currently being executed or finished running and cannot be cancelled.
True if the call can be canceled.

Return type bool

done (*self*)

Returns Whether the computation is done.

Return type bool

get_result (*self*)

Gets the computation result. Will block until the computation is finished.

Raises

- **Exception** – If computation fails. Returns traceback.
- **cf.TimeoutError** – If computation takes longer than default timeout time.
- **cf.CancelledError** – If computation was canceled before completing.

Returns The result of the requested job.

property is_cancelled (*self*)

Returns Returns whether computation was cancelled.

Return type bool

class evalml.automl.engine.cf_engine.**CFEngine** (*client=None*)

The concurrent.futures (CF) engine

Parameters **client** (*None* or **CFClient**) – If *None*, creates a threaded pool for processing.
Defaults to *None*.

Methods

<i>close</i>	Function to properly shutdown the Engine's Client's resources.
<i>is_closed</i>	Property that determines whether the Engine's Client's resources are shutdown.
<i>setup_job_log</i>	
<i>submit_evaluation_job</i>	Send evaluation job to cluster.

continues on next page

Table 48 – continued from previous page

<code>submit_scoring_job</code>	Send scoring job to cluster.
<code>submit_training_job</code>	Send training job to cluster.

close (*self*)

Function to properly shutdown the Engine’s Client’s resources.

property is_closed (*self*)

Property that determines whether the Engine’s Client’s resources are shutdown.

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*) →
evalml.automl.engine.engine_base.EngineComputation

Send evaluation job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to evaluate
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

an object wrapping a reference to a future-like computation occurring in the resource pool

Return type *CFComputation*

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*) →
evalml.automl.engine.engine_base.EngineComputation

Send scoring job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the resource pool

Return type *CFComputation*

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*) →
evalml.automl.engine.engine_base.EngineComputation

Send training job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

an object wrapping a reference to a future-like computation occurring in the resource pool

Return type *CFComputation*

dask_engine**Module Contents****Classes Summary**

<i>DaskComputation</i>	A Future-like wrapper around jobs created by the DaskEngine.
<i>DaskEngine</i>	The dask engine

Contents

class evalml automl engine dask_engine **DaskComputation** (*dask_future*)

A Future-like wrapper around jobs created by the DaskEngine.

Parameters **dask_future** (*callable*) – Computation to do.

Methods

<i>cancel</i>	Cancel the current computation.
<i>done</i>	returns Whether the computation is done.
<i>get_result</i>	Gets the computation result.
<i>is_cancelled</i>	returns Returns whether computation was cancelled.

cancel (*self*)

Cancel the current computation.

done (*self*)

Returns Whether the computation is done.

Return type bool

get_result (*self*)

Gets the computation result. Will block until the computation is finished.

Raises **Exception** – If computation fails. Returns traceback.

property **is_cancelled** (*self*)

Returns Returns whether computation was cancelled.

Return type bool

class evalml.automl.engine.dask_engine.**DaskEngine** (*cluster=None*)

The dask engine

Parameters **cluster** (*None or dd.Client*) – If None, creates a local, threaded Dask client for processing. Defaults to None.

Methods

<code>close</code>	Closes the underlying cluster.
<code>is_closed</code>	Property that determines whether the Engine's Client's resources are shutdown.
<code>send_data_to_cluster</code>	Send data to the cluster.
<code>setup_job_log</code>	
<code>submit_evaluation_job</code>	Send evaluation job to cluster.
<code>submit_scoring_job</code>	Send scoring job to cluster.
<code>submit_training_job</code>	Send training job to cluster.

close (*self*)

Closes the underlying cluster.

property is_closed (*self*)

Property that determines whether the Engine's Client's resources are shutdown.

send_data_to_cluster (*self, X, y*)

Send data to the cluster.

The implementation uses caching so the data is only sent once. This follows dask best practices.

Parameters

- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns the modeling data

Return type dask.Future

static setup_job_log ()

submit_evaluation_job (*self, automl_config, pipeline, X, y*) → *evalml.automl.engine.engine_base.EngineComputation*

Send evaluation job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to evaluate
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the dask cluster

Return type *DaskComputation*

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*) →
evalml.automl.engine.engine_base.EngineComputation

Send scoring job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the dask cluster

Return type *DaskComputation*

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*) →
evalml.automl.engine.engine_base.EngineComputation

Send training job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the dask cluster

Return type *DaskComputation*

engine_base

Module Contents

Classes Summary

<i>EngineBase</i>	Helper class that provides a standard way to create an ABC using
<i>EngineComputation</i>	Wrapper around the result of a (possibly asynchronous) engine computation.
<i>JobLogger</i>	Mimics the behavior of a python logging.Logger but stores all messages rather than actually logging them.

Functions

<code>evaluate_pipeline</code>	Function submitted to the <code>submit_evaluation_job</code> engine method.
<code>score_pipeline</code>	Wrapper around <code>pipeline.score</code> method to make it easy to score pipelines with dask.
<code>train_and_score_pipeline</code>	Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores
<code>train_pipeline</code>	Train a pipeline and tune the threshold if necessary.

Contents

class `evalml.engine.engine_base.EngineBase`

Helper class that provides a standard way to create an ABC using inheritance.

Methods

<code>setup_job_log</code>	
<code>submit_evaluation_job</code>	Submit job for pipeline evaluation during AutoMLSearch.
<code>submit_scoring_job</code>	Submit job for pipeline scoring.
<code>submit_training_job</code>	Submit job for pipeline training.

static `setup_job_log()`

abstract `submit_evaluation_job(self, automl_config, pipeline, X, y)`

Submit job for pipeline evaluation during AutoMLSearch.

abstract `submit_scoring_job(self, automl_config, pipeline, X, y, objectives)`

Submit job for pipeline scoring.

abstract `submit_training_job(self, automl_config, pipeline, X, y)`

Submit job for pipeline training.

class `evalml.engine.engine_base.EngineComputation`

Wrapper around the result of a (possibly asynchronous) engine computation.

Methods

<code>cancel</code>	Cancel the computation.
<code>done</code>	Whether the computation is done.
<code>get_result</code>	Gets the computation result.

abstract `cancel(self)`

Cancel the computation.

abstract `done(self)`

Whether the computation is done.

abstract `get_result(self)`

Gets the computation result. Will block until the computation is finished.

Raises Exception: If computation fails. Returns traceback.

`evalml automl engine engine_base.evaluate_pipeline` (*pipeline*, *automl_config*, *X*, *y*, *logger*)

Function submitted to the `submit_evaluation_job` engine method.

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to score
- **automl_config** (*AutoMLConfig*) – The `AutoMLSearch` object, used to access config and the error callback
- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Training target

Returns

First - A dict containing `cv_score_mean`, `cv_scores`, `training_time` and a `cv_data` structure with details.
Second - The pipeline class we trained and scored. Third - the job logger instance with all the recorded messages.

Return type tuple of three items

class `evalml automl engine engine_base.JobLogger`

Mimics the behavior of a python logging.Logger but stores all messages rather than actually logging them.

This is used during engine jobs so that log messages are recorded after the job completes. This is desired so that all of the messages for a single job are grouped together in the log.

Methods

<code>debug</code>	Store message at the debug level.
<code>error</code>	Store message at the error level.
<code>info</code>	Store message at the info level.
<code>warning</code>	Store message at the warning level.
<code>write_to_logger</code>	Write all the messages to the logger. First In First Out order.

debug (*self*, *msg*)
Store message at the debug level.

error (*self*, *msg*)
Store message at the error level.

info (*self*, *msg*)
Store message at the info level.

warning (*self*, *msg*)
Store message at the warning level.

write_to_logger (*self*, *logger*)
Write all the messages to the logger. First In First Out order.

`evalml automl engine engine_base.score_pipeline` (*pipeline*, *X*, *y*, *objectives*, *X_schema=None*, *y_schema=None*)

Wrapper around `pipeline.score` method to make it easy to score pipelines with dask.

Arguments: `pipeline` (*PipelineBase*): The pipeline to score. `X` (*pd.DataFrame*): Features to score on. `y` (*pd.Series*): Target used to calculate scores. `X_schema` (*ww.TableSchema*): Schema for features. `y_schema` (*ww.ColumnSchema*): Schema for columns.

Returns dict containing pipeline scores.

```
evalml.automl.engine.engine_base.train_and_score_pipeline(pipeline, automl_config,  
                                                         full_X_train,  
                                                         full_y_train, logger)
```

Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to score
- **automl_config** (*AutoMLSearch*) – The AutoMLSearch object, used to access config and the error callback
- **full_X_train** (*pd.DataFrame*) – Training features
- **full_y_train** (*pd.Series*) – Training target

Returns

First - A dict containing cv_score_mean, cv_scores, training_time and a cv_data structure with details.
Second - The pipeline class we trained and scored. Third - the job logger instance with all the recorded messages.

Return type tuple of three items

```
evalml.automl.engine.engine_base.train_pipeline(pipeline, X, y, automl_config,  
                                                schema=True)
```

Train a pipeline and tune the threshold if necessary.

Parameters

- **pipeline** (*PipelineBase*) – Pipeline to train.
- **x** (*pd.DataFrame*) – Features to train on.
- **y** (*pd.Series*) – Target to train on.
- **automl_config** (*AutoMLSearch*) – The AutoMLSearch object, used to access config and the error callback
- **schema** (*bool*) – Whether to use the schemas for X and y

Returns trained pipeline.

Return type pipeline (*PipelineBase*)

sequential_engine

Module Contents

Classes Summary

<i>SequentialComputation</i>	A Future-like api for jobs created by the SequentialEngine, an Engine that sequentially computes the submitted jobs.
<i>SequentialEngine</i>	The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Contents

class evalml.engine.sequential_engine.**SequentialComputation** (*work*,
***kwargs*)

A Future-like api for jobs created by the SequentialEngine, an Engine that sequentially computes the submitted jobs.

In order to separate the engine from the AutoMLSearch loop, we need the sequential computations to behave the same way as concurrent computations from AutoMLSearch’s point-of-view. One way to do this is by delaying the computation in the sequential engine until `get_result` is called. Since AutoMLSearch will call `get_result` only when the computation is “done”, by always returning `True` in `done()` we make sure that `get_result` is called in the order that the jobs are submitted. So the computations happen sequentially!

Parameters *work* (*callable*) – Computation that should be done by the engine.

Methods

<code>cancel</code>	Cancel the current computation.
<code>done</code>	Whether the computation is done.
<code>get_result</code>	Gets the computation result.

cancel (*self*)

Cancel the current computation.

done (*self*)

Whether the computation is done.

get_result (*self*)

Gets the computation result. Will block until the computation is finished.

Raises Exception: If computation fails. Returns traceback.

class evalml.engine.sequential_engine.**SequentialEngine**

The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Methods

<code>close</code>	
<code>setup_job_log</code>	
<code>submit_evaluation_job</code>	Submit job for pipeline evaluation during AutoMLSearch.
<code>submit_scoring_job</code>	Submit job for pipeline scoring.
<code>submit_training_job</code>	Submit job for pipeline training.

close (*self*)

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline evaluation during AutoMLSearch.

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*)

Submit job for pipeline scoring.

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline training.

Package Contents

Classes Summary

<code>CFEngine</code>	The concurrent.futures (CF) engine
<code>DaskEngine</code>	The dask engine
<code>EngineBase</code>	Helper class that provides a standard way to create an ABC using
<code>EngineComputation</code>	Wrapper around the result of a (possibly asynchronous) engine computation.
<code>SequentialEngine</code>	The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Functions

<code>evaluate_pipeline</code>	Function submitted to the <code>submit_evaluation_job</code> engine method.
<code>train_and_score_pipeline</code>	Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores
<code>train_pipeline</code>	Train a pipeline and tune the threshold if necessary.

Contents

class `evalml.automl.engine.CFEngine` (*client=None*)

The concurrent.futures (CF) engine

Parameters **client** (*None* or *CFClient*) – If *None*, creates a threaded pool for processing.
Defaults to *None*.

Methods

<code>close</code>	Function to properly shutdown the Engine's Client's resources.
<code>is_closed</code>	Property that determines whether the Engine's Client's resources are shutdown.
<code>setup_job_log</code>	
<code>submit_evaluation_job</code>	Send evaluation job to cluster.
<code>submit_scoring_job</code>	Send scoring job to cluster.
<code>submit_training_job</code>	Send training job to cluster.

close (*self*)

Function to properly shutdown the Engine's Client's resources.

property is_closed (*self*)

Property that determines whether the Engine's Client's resources are shutdown.

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*) →
evalml.automl.engine.engine_base.EngineComputation

Send evaluation job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to evaluate
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

an object wrapping a reference to a future-like computation occurring in the resource pool

Return type CFComputation

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*) → *evalml.automl.engine.engine_base.EngineComputation*

Send scoring job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the resource pool

Return type CFComputation

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*) → *evalml.automl.engine.engine_base.EngineComputation*

Send training job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

an object wrapping a reference to a future-like computation occurring in the resource pool

Return type CFComputation

class *evalml.automl.engine.DaskEngine* (*cluster=None*)

The dask engine

Parameters **cluster** (*None* or *dd.Client*) – If *None*, creates a local, threaded Dask client for processing. Defaults to *None*.

Methods

<code>close</code>	Closes the underlying cluster.
<code>is_closed</code>	Property that determines whether the Engine's Client's resources are shutdown.
<code>send_data_to_cluster</code>	Send data to the cluster.
<code>setup_job_log</code>	
<code>submit_evaluation_job</code>	Send evaluation job to cluster.
<code>submit_scoring_job</code>	Send scoring job to cluster.
<code>submit_training_job</code>	Send training job to cluster.

close (*self*)

Closes the underlying cluster.

property is_closed (*self*)

Property that determines whether the Engine's Client's resources are shutdown.

send_data_to_cluster (*self*, *X*, *y*)

Send data to the cluster.

The implementation uses caching so the data is only sent once. This follows dask best practices.

Parameters

- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns the modeling data

Return type *dask.Future*

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*) →
evalml.automl.engine.engine_base.EngineComputation

Send evaluation job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to evaluate
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the dask cluster

Return type *DaskComputation*

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*) →
evalml.automl.engine.engine_base.EngineComputation

Send scoring job to cluster.

Parameters

- **automl_config** – structure containing data passed from AutoMLSearch instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train

- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the `dask` cluster

Return type `DaskComputation`

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*) → *evalml.automl.engine.engine_base.EngineComputation*

Send training job to cluster.

Parameters

- **automl_config** – structure containing data passed from `AutoMLSearch` instance
- **pipeline** (*pipeline.PipelineBase*) – pipeline to train
- **X** (*pd.DataFrame*) – input data for modeling
- **y** (*pd.Series*) – target data for modeling

Returns

a object wrapping a reference to a future-like computation occurring in the `dask` cluster

Return type `DaskComputation`

class `evalml.automl.engine.EngineBase`

Helper class that provides a standard way to create an ABC using inheritance.

Methods

setup_job_log

<i>submit_evaluation_job</i>	Submit job for pipeline evaluation during <code>AutoMLSearch</code> .
------------------------------	---

<i>submit_scoring_job</i>	Submit job for pipeline scoring.
---------------------------	----------------------------------

<i>submit_training_job</i>	Submit job for pipeline training.
----------------------------	-----------------------------------

static `setup_job_log()`

abstract `submit_evaluation_job` (*self*, *automl_config*, *pipeline*, *X*, *y*)
Submit job for pipeline evaluation during `AutoMLSearch`.

abstract `submit_scoring_job` (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*)
Submit job for pipeline scoring.

abstract `submit_training_job` (*self*, *automl_config*, *pipeline*, *X*, *y*)
Submit job for pipeline training.

class `evalml.automl.engine.EngineComputation`

Wrapper around the result of a (possibly asynchronous) engine computation.

Methods

<i>cancel</i>	Cancel the computation.
---------------	-------------------------

<i>done</i>	Whether the computation is done.
-------------	----------------------------------

<i>get_result</i>	Gets the computation result.
-------------------	------------------------------

abstract `cancel` (*self*)

Cancel the computation.

abstract done (*self*)

Whether the computation is done.

abstract get_result (*self*)

Gets the computation result. Will block until the computation is finished.

Raises Exception: If computation fails. Returns traceback.

`evalml automl engine evaluate_pipeline (pipeline, automl_config, X, y, logger)`

Function submitted to the submit_evaluation_job engine method.

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to score
- **automl_config** (*AutoMLConfig*) – The AutoMLSearch object, used to access config and the error callback
- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Training target

Returns

First - A dict containing cv_score_mean, cv_scores, training_time and a cv_data structure with details.

Second - The pipeline class we trained and scored. Third - the job logger instance with all the recorded messages.

Return type tuple of three items

class `evalml automl engine SequentialEngine`

The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Methods

close

setup_job_log

<i>submit_evaluation_job</i>	Submit job for pipeline evaluation during AutoMLSearch.
------------------------------	---

<i>submit_scoring_job</i>	Submit job for pipeline scoring.
---------------------------	----------------------------------

<i>submit_training_job</i>	Submit job for pipeline training.
----------------------------	-----------------------------------

close (*self*)

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline evaluation during AutoMLSearch.

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*)

Submit job for pipeline scoring.

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline training.

`evalml automl engine train_and_score_pipeline (pipeline, automl_config, full_X_train, full_y_train, logger)`

Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to score
- **automl_config** (*AutoMLSearch*) – The AutoMLSearch object, used to access config and the error callback
- **full_x_train** (*pd.DataFrame*) – Training features
- **full_y_train** (*pd.Series*) – Training target

Returns

First - A dict containing **cv_score_mean**, **cv_scores**, **training_time** and a **cv_data** structure with details.
Second - The pipeline class we trained and scored. **Third** - the job logger instance with all the recorded messages.

Return type tuple of three items

`evalml.automl.engine.train_pipeline(pipeline, X, y, automl_config, schema=True)`

Train a pipeline and tune the threshold if necessary.

Parameters

- **pipeline** (*PipelineBase*) – Pipeline to train.
- **x** (*pd.DataFrame*) – Features to train on.
- **y** (*pd.Series*) – Target to train on.
- **automl_config** (*AutoMLSearch*) – The AutoMLSearch object, used to access config and the error callback
- **schema** (*bool*) – Whether to use the schemas for X and y

Returns trained pipeline.

Return type pipeline (*PipelineBase*)

Submodules**automl_search****Module Contents****Classes Summary**

<i>AutoMLSearch</i>	Automated Pipeline search.
---------------------	----------------------------

Functions

<i>build_engine_from_str</i>	Function that converts a convenience string for an parallel engine type and returns an instance of that engine.
<i>search</i>	Given data and configuration, run an automl search.
<i>search_iterative</i>	Given data and configuration, run an automl search.

Contents

```
class evalml.automl.automl_search.AutoMLSearch(X_train=None, y_train=None,
                                              problem_type=None, objective='auto',
                                              max_iterations=None, max_time=None,
                                              patience=None, tolerance=None,
                                              data_splitter=None, allowed_component_graphs=None,
                                              allowed_model_families=None,
                                              start_iteration_callback=None,
                                              add_result_callback=None,
                                              error_callback=None,
                                              additional_objectives=None,
                                              alternate_thresholding_objective='F1',
                                              random_seed=0, n_jobs=-1,
                                              tuner_class=None, optimize_thresholds=True,
                                              ensembling=False, max_batches=None,
                                              problem_configuration=None,
                                              train_best_pipeline=True,
                                              pipeline_parameters=None,
                                              custom_hyperparameters=None,
                                              sampler_method='auto',
                                              sampler_balanced_ratio=0.25,
                                              _ensembling_split_size=0.2,
                                              _pipelines_per_batch=5,
                                              _automl_algorithm='iterative',
                                              engine='sequential', verbose=False)
```

Automated Pipeline search.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.
- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.
- **max_iterations** (*int*) – Maximum number of iterations to search. If `max_iterations` and `max_time` is not set, then `max_iterations` will default to `max_iterations` of 5.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in seconds. For strings, time can be specified as seconds, minutes, or hours.

- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If None, early stopping is disabled. Defaults to None.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if patience is not None. Defaults to None.
- **allowed_component_graphs** (*dict*) – A dictionary of lists or ComponentGraphs indicating the component graphs allowed in the search. The format should follow { “Name_0”: [list_of_components], “Name_1”: [ComponentGraph(...)] }

The default of None indicates all pipeline component graphs for this problem type are allowed. Setting this field will cause allowed_model_families to be ignored.

e.g. allowed_component_graphs = { “My_Graph”: [“Imputer”, “One Hot Encoder”, “Random Forest Classifier”] }

- **allowed_model_families** (*list(str, ModelFamily)*) – The model families to search. The default of None searches over all model families. Run evalml.pipelines.components.utils.allowed_model_families(“binary”) to see options. Change *binary* to *multiclass* or *regression* depending on the problem type. Note that if allowed_pipelines is provided, this parameter will be ignored.
- **data_splitter** (*sklearn.model_selection.BaseCrossValidator*) – Data splitting method to use. Defaults to StratifiedKFold.
- **tuner_class** – The tuner class to use. Defaults to SKOptTuner.
- **optimize_thresholds** (*bool*) – Whether or not to optimize the binary pipeline threshold. Defaults to True.
- **start_iteration_callback** (*callable*) – Function called before each pipeline training iteration. Callback function takes three positional parameters: The pipeline instance and the AutoMLSearch object.
- **add_result_callback** (*callable*) – Function called after each pipeline training iteration. Callback function takes three positional parameters: A dictionary containing the training results for the new pipeline, an untrained_pipeline containing the parameters used during training, and the AutoMLSearch object.
- **error_callback** (*callable*) – Function called when *search()* errors and raises an Exception. Callback function takes three positional parameters: the Exception raised, the traceback, and the AutoMLSearch object. Must also accepts kwargs, so AutoMLSearch is able to pass along other appropriate parameters by default. Defaults to None, which will call *log_error_callback*.
- **additional_objectives** (*list*) – Custom set of objectives to score on. Will override default objectives for problem type if not empty.
- **alternate_thresholding_objective** (*str*) – The objective to use for thresholding binary classification pipelines if the main objective provided isn’t tuneable. Defaults to F1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used.
- **ensembling** (*boolean*) – If True, runs ensembling in a separate batch after every allowed pipeline class has been iterated over. If the number of unique pipelines to search over per batch is one, ensembling will not run. Defaults to False.

- **max_batches** (*int*) – The maximum number of batches of pipelines to search. Parameters `max_time`, and `max_iterations` have precedence over stopping the search.
- **problem_configuration** (*dict*, *None*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, `forecast_horizon`, and `max_delay` variables.
- **train_best_pipeline** (*boolean*) – Whether or not to train the best pipeline before returning it. Defaults to `True`.
- **pipeline_parameters** (*dict*) – A dict of the parameters used to initialize a pipeline with. Keys should consist of the component names and values should specify parameter values
e.g. `pipeline_parameters = { 'Imputer' : { 'numeric_impute_strategy': 'most_frequent' } }`
- **custom_hyperparameters** (*dict*) – A dict of the hyperparameter ranges used to iterate over during search. Keys should consist of the component names and values should specify a singular value or `skopt.Space`.
e.g. `custom_hyperparameters = { 'Imputer' : { 'numeric_impute_strategy': Categorical(['most_frequent', 'median']) } }`
- **sampler_method** (*str*) – The data sampling component to use in the pipelines if the problem type is classification and the target balance is smaller than the `sampler_balanced_ratio`. Either `'auto'`, which will use our preferred sampler for the data, `'Undersampler'`, `'Oversampler'`, or `None`. Defaults to `'auto'`.
- **sampler_balanced_ratio** (*float*) – The minority:majority class ratio that we consider balanced, so a 1:4 ratio would be equal to 0.25. If the class balance is larger than this provided value, then we will not add a sampler since the data is then considered balanced. Overrides the `sampler_ratio` of the samplers. Defaults to 0.25.
- **_ensembling_split_size** (*float*) – The amount of the training data we'll set aside for training ensemble metalearners. Only used when `ensembling` is `True`. Must be between 0 and 1, exclusive. Defaults to 0.2
- **pipelines_per_batch** (*int*) – The number of pipelines to train for every batch after the first one. The first batch will train a baseline pipeline + one of each pipeline family allowed in the search.
- **_automl_algorithm** (*str*) – The automl algorithm to use. Currently the two choices are `'iterative'` and `'default'`. Defaults to `iterative`.
- **engine** (*EngineBase* or *str*) – The engine instance used to evaluate pipelines. Dask or `concurrent.futures` engines can also be chosen by providing a string from the list [`"sequential"`, `"cf_threaded"`, `"cf_process"`, `"dask_threaded"`, `"dask_process"`]. If a parallel engine is selected this way, the maximum amount of parallelism, as determined by the engine, will be used. Defaults to `"sequential"`.
- **verbose** (*boolean*) – Whether or not to display semi-real-time updates to stdout while search is running. Defaults to `False`.

Methods

`add_to_rankings`

Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.

continues on next page

Table 69 – continued from previous page

<code>best_pipeline</code>	Returns a trained instance of the best pipeline and parameters found during automl search. If <code>train_best_pipeline</code> is set to False, returns an untrained pipeline instance.
<code>close_engine</code>	Function to explicitly close the engine, client, parallel resources.
<code>describe_pipeline</code>	Describe a pipeline
<code>full_rankings</code>	Returns a pandas.DataFrame with scoring results from all pipelines searched
<code>get_pipeline</code>	Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline
<code>load</code>	Loads AutoML object at file path
<code>plot</code>	
<code>rankings</code>	Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.
<code>results</code>	Class that allows access to a copy of the results from <code>automl_search</code> .
<code>save</code>	Saves AutoML object at file path
<code>score_pipelines</code>	Score a list of pipelines on the given holdout data.
<code>search</code>	Find the best pipeline for the data set.
<code>train_pipelines</code>	Train a list of pipelines on the training data.

add_to_rankings (*self*, *pipeline*)

Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.

Parameters *pipeline* (*PipelineBase*) – pipeline to train and evaluate.

property best_pipeline (*self*)

Returns a trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Returns A trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Return type *PipelineBase*

close_engine (*self*)

Function to explicitly close the engine, client, parallel resources.

describe_pipeline (*self*, *pipeline_id*, *return_dict=False*)

Describe a pipeline

Parameters

- **pipeline_id** (*int*) – pipeline to describe
- **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Description of specified pipeline. Includes information such as type of pipeline components, problem, training time, cross validation, etc.

property full_rankings (*self*)

Returns a pandas.DataFrame with scoring results from all pipelines searched

get_pipeline (*self*, *pipeline_id*)

Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline initialized with the parameters used to train that pipeline during automl search.

Parameters **pipeline_id** (*int*) – pipeline to retrieve

Returns untrained pipeline instance associated with the provided ID

Return type PipelineBase

static load (*file_path*, *pickle_type*='cloudpickle')

Loads AutoML object at file path

Parameters

- **file_path** (*str*) – location to find file to load
- {"**pickle**" (*pickle_type*)} – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.
- "**cloudpickle**" – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.

Returns AutoSearchBase object

property plot (*self*)

property rankings (*self*)

Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.

property results (*self*)

Class that allows access to a copy of the results from *automl_search*.

Returns: dict containing **pipeline_results**: a dict with results from each pipeline, and **search_order**: a list describing the order the pipelines were searched.

save (*self*, *file_path*, *pickle_type*='cloudpickle', *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves AutoML object at file path

Parameters

- **file_path** (*str*) – location to save file
- {"**pickle**" (*pickle_type*)} – the pickling library to use.
- "**cloudpickle**" – the pickling library to use.
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score_pipelines (*self*, *pipelines*, *X_holdout*, *y_holdout*, *objectives*)

Score a list of pipelines on the given holdout data.

Parameters

- **pipelines** (*list* (PipelineBase)) – List of pipelines to train.
- **X_holdout** (*pd.DataFrame*) – Holdout features.
- **y_holdout** (*pd.Series*) – Holdout targets for scoring.
- **objectives** (*list* (*str*), *list* (ObjectiveBase)) – Objectives used for scoring.

Returns Dictionary keyed by pipeline name that maps to a dictionary of scores. Note that the any pipelines that error out during scoring will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, Dict[str, float]]

search (*self*, *show_iteration_plot=True*)

Find the best pipeline for the data set.

Parameters

- **feature_types** (*list*, *optional*) – list of feature types, either numerical or categorical. Categorical features will automatically be encoded
- **show_iteration_plot** (*boolean*, *True*) – Shows an iteration vs. score plot in Jupyter notebook. Disabled by default in non-Jupyter environments.

train_pipelines (*self*, *pipelines*)

Train a list of pipelines on the training data.

This can be helpful for training pipelines once the search is complete.

Parameters **pipelines** (*list* (*PipelineBase*)) – List of pipelines to train.

Returns Dictionary keyed by pipeline name that maps to the fitted pipeline. Note that the any pipelines that error out during training will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, PipelineBase]

`evalml.automl.automl_search.build_engine_from_str(engine_str)`

Function that converts a convenience string for an parallel engine type and returns an instance of that engine.

Parameters **engine_str** (*str*) – String representing the requested engine.

Returns Instance of the requested engine.

Return type (EngineBase)

`evalml.automl.automl_search.search(X_train=None, y_train=None, problem_type=None, objective='auto', mode='fast', max_time=None, patience=None, tolerance=None, problem_configuration=None)`

Given data and configuration, run an automl search.

This method will run EvalML's default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you'd like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.
- **mode** (*str*) – mode for DefaultAlgorithm. There are two modes: fast and long, where fast is a subset of long. Please look at DefaultAlgorithm for more details.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in seconds. For strings, time can be specified as seconds, minutes, or hours.
- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If None, early stopping is disabled. Defaults to None.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if patience is not None. Defaults to None.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the date_index, gap, and max_delay variables.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

```
evalml.automl.automl_search.search_iterative(X_train=None, y_train=None, problem_type=None, objective='auto', problem_configuration=None, **kwargs)
```

Given data and configuration, run an automl search.

This method will run EvalML’s default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you’d like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See evalml.problem_types.ProblemType.all_problem_types for a full list.
- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and

- R2 for regression problems.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example,
- **time series problems** (*in*) –
- **should be passed in for the date_index** (*values*) –
- **gap** –
- **forecast_horizon** –
- **max_delay variables.** (*and*) –

Other keyword arguments which are provided will be passed to AutoMLSearch.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

callbacks

Module Contents

Functions

<i>log_error_callback</i>	Logs the exception thrown as an error. Will not throw. This is the default behavior for AutoMLSearch.
<i>raise_error_callback</i>	Raises the exception thrown by the AutoMLSearch object. Also logs the exception as an error.
<i>silent_error_callback</i>	No-op.

Attributes Summary

<i>logger</i>

Contents

`evalml.automl.callbacks.log_error_callback(exception, traceback, automl, **kwargs)`

Logs the exception thrown as an error. Will not throw. This is the default behavior for AutoMLSearch.

`evalml.automl.callbacks.logger`

`evalml.automl.callbacks.raise_error_callback(exception, traceback, automl, **kwargs)`

Raises the exception thrown by the AutoMLSearch object. Also logs the exception as an error.

`evalml.automl.callbacks.silent_error_callback(exception, traceback, automl, **kwargs)`

No-op.

pipeline_search_plots

Module Contents

Classes Summary

<code>PipelineSearchPlots</code>	Plots for the AutoMLSearch class.
<code>SearchIterationPlot</code>	

Contents

class evalml automl.pipeline_search_plots.**PipelineSearchPlots** (*results*, *objective*)

Plots for the AutoMLSearch class.

Methods

<code>search_iteration_plot</code>	Shows a plot of the best score at each iteration using data gathered during training.
------------------------------------	---

search_iteration_plot (*self*, *interactive_plot=False*)

Shows a plot of the best score at each iteration using data gathered during training.

Returns plot

class evalml automl.pipeline_search_plots.**SearchIterationPlot** (*results*, *objective*)

Methods

<code>update</code>	
---------------------	--

update (*self*, *results*, *objective*)

utils

Module Contents

Functions

<code>check_all_pipeline_names_unique</code>	Checks whether all the pipeline names are unique.
<code>get_best_sampler_for_data</code>	Returns the name of the sampler component to use for AutoMLSearch.
<code>get_default_primary_search_objective</code>	Get the default primary search objective for a problem type.
<code>get_pipelines_from_component_graphs</code>	Returns created pipelines from passed component graphs based on the specified problem type.

continues on next page

Table 75 – continued from previous page

<code>make_data_splitter</code>	Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.
<code>tune_binary_threshold</code>	Tunes the threshold of a binary pipeline to the X and y thresholding data

Attributes Summary

`AutoMLConfig`

Contents

`evalml automl utils AutoMLConfig`

`evalml automl utils check_all_pipeline_names_unique (pipelines)`

Checks whether all the pipeline names are unique.

Parameters `pipelines` (*list (PipelineBase)*) – List of pipelines to check if all names are unique.

Returns None

Raises `ValueError` – if any pipeline names are duplicated.

`evalml automl utils get_best_sampler_for_data (X, y, sampler_method, sampler_balanced_ratio)`

Returns the name of the sampler component to use for AutoMLSearch.

Parameters

- `X` (*pd.DataFrame*) – The input feature data
- `y` (*pd.Series*) – The input target data
- `sampler_method` (*str*) – The `sampler_type` argument passed to AutoMLSearch
- `sampler_balanced_ratio` (*float*) – The ratio of min:majority targets that we would consider balanced, or should balance the classes to.

Returns The string name of the sampling component to use, or None if no sampler is necessary

Return type `str`, None

`evalml automl utils get_default_primary_search_objective (problem_type)`

Get the default primary search objective for a problem type.

Parameters `problem_type` (*str or ProblemType*) – problem type of interest.

Returns primary objective instance for the problem type.

Return type `ObjectiveBase`

`evalml automl utils get_pipelines_from_component_graphs (component_graphs_dict, problem_type, parameters=None, random_seed=0)`

Returns created pipelines from passed component graphs based on the specified problem type.

Parameters

- **component_graphs_dict** (*dict*) – The dict of component graphs.
- **problem_type** (*str or ProblemType*) – The problem type for which pipelines will be created.
- **parameters** (*dict or None*) – Pipeline-level parameters that should be passed to the proposed pipelines.
- **random_seed** (*int*) – Random seed.

Returns List of pipelines made from the passed component graphs.

Return type list

```
evalml automl utils make_data_splitter (X, y, problem_type, problem_configuration=None,  
                                         n_splits=3, shuffle=True, random_seed=0)
```

Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.

Parameters

- **X** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].
- **problem_type** (*ProblemType*) – The type of machine learning problem.
- **problem_configuration** (*dict, None*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the date_index, gap, and max_delay variables. Defaults to None.
- **n_splits** (*int, None*) – The number of CV splits, if applicable. Defaults to 3.
- **shuffle** (*bool*) – Whether or not to shuffle the data before splitting, if applicable. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Data splitting method.

Return type sklearn.model_selection.BaseCrossValidator

```
evalml automl utils tune_binary_threshold (pipeline, objective, problem_type,  
                                           X_threshold_tuning, y_threshold_tuning)
```

Tunes the threshold of a binary pipeline to the X and y thresholding data

Parameters

- **pipeline** (*Pipeline*) – Pipeline instance to threshold.
- **objective** (*ObjectiveBase*) – The objective we want to tune with. If not tuneable and best_pipeline is True, will use F1.
- **problem_type** (*ProblemType*) – The problem type of the pipeline.
- **X_threshold_tuning** (*pd.DataFrame*) – Features to tune pipeline to.
- **y_threshold_tuning** (*pd.Series*) – Target data to tune pipeline to.

Package Contents

Classes Summary

<i>AutoMLSearch</i>	Automated Pipeline search.
<i>EngineBase</i>	Helper class that provides a standard way to create an ABC using
<i>SequentialEngine</i>	The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Functions

<i>get_default_primary_search_objective</i>	Get the default primary search objective for a problem type.
<i>make_data_splitter</i>	Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.
<i>search</i>	Given data and configuration, run an automl search.
<i>search_iterative</i>	Given data and configuration, run an automl search.
<i>tune_binary_threshold</i>	Tunes the threshold of a binary pipeline to the X and y thresholding data

Contents

```
class evalml.automl.AutoMLSearch(X_train=None, y_train=None, problem_type=None,
                                  objective='auto', max_iterations=None,
                                  max_time=None, patience=None, tolerance=None,
                                  data_splitter=None, allowed_component_graphs=None,
                                  allowed_model_families=None,
                                  start_iteration_callback=None, add_result_callback=None,
                                  error_callback=None, additional_objectives=None, al-
                                  ternate_thresholding_objective='F1', random_seed=0,
                                  n_jobs=- 1, tuner_class=None, optimize_thresholds=True,
                                  ensembling=False, max_batches=None, prob-
                                  lem_configuration=None, train_best_pipeline=True,
                                  pipeline_parameters=None, custom_hyperparameters=None,
                                  sampler_method='auto', sampler_balanced_ratio=0.25,
                                  _ensembling_split_size=0.2, _pipelines_per_batch=5,
                                  _automl_algorithm='iterative', engine='sequential', ver-
                                  bose=False)
```

Automated Pipeline search.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem.

See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - `LogLossBinary` for binary classification problems,
 - `LogLossMulticlass` for multiclass classification problems, and
 - `R2` for regression problems.
- **max_iterations** (*int*) – Maximum number of iterations to search. If `max_iterations` and `max_time` is not set, then `max_iterations` will default to `max_iterations` of 5.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in seconds. For strings, time can be specified as seconds, minutes, or hours.
- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If `None`, early stopping is disabled. Defaults to `None`.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if `patience` is not `None`. Defaults to `None`.
- **allowed_component_graphs** (*dict*) – A dictionary of lists or `ComponentGraphs` indicating the component graphs allowed in the search. The format should follow { “Name_0”: [list_of_components], “Name_1”: [ComponentGraph(...)] }

The default of `None` indicates all pipeline component graphs for this problem type are allowed. Setting this field will cause `allowed_model_families` to be ignored.

e.g. `allowed_component_graphs = { “My_Graph”: [“Imputer”, “One Hot Encoder”, “Random Forest Classifier”] }`

- **allowed_model_families** (*list(str, ModelFamily)*) – The model families to search. The default of `None` searches over all model families. Run `evalml.pipelines.components.utils.allowed_model_families(“binary”)` to see options. Change *binary* to *multiclass* or *regression* depending on the problem type. Note that if `allowed_pipelines` is provided, this parameter will be ignored.
- **data_splitter** (*sklearn.model_selection.BaseCrossValidator*) – Data splitting method to use. Defaults to `StratifiedKFold`.
- **tuner_class** – The tuner class to use. Defaults to `SKOptTuner`.
- **optimize_thresholds** (*bool*) – Whether or not to optimize the binary pipeline threshold. Defaults to `True`.
- **start_iteration_callback** (*callable*) – Function called before each pipeline training iteration. Callback function takes three positional parameters: The pipeline instance and the `AutoMLSearch` object.
- **add_result_callback** (*callable*) – Function called after each pipeline training iteration. Callback function takes three positional parameters: A dictionary containing the training results for the new pipeline, an `untrained_pipeline` containing the parameters used during training, and the `AutoMLSearch` object.
- **error_callback** (*callable*) – Function called when `search()` errors and raises an `Exception`. Callback function takes three positional parameters: the `Exception` raised, the `traceback`, and the `AutoMLSearch` object. Must also accepts `kwargs`, so `AutoMLSearch` is able to pass along other appropriate parameters by default. Defaults to `None`, which will call `log_error_callback`.

- **additional_objectives** (*list*) – Custom set of objectives to score on. Will override default objectives for problem type if not empty.
- **alternate_thresholding_objective** (*str*) – The objective to use for thresholding binary classification pipelines if the main objective provided isn't tuneable. Defaults to F1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used.
- **ensembling** (*boolean*) – If True, runs ensembling in a separate batch after every allowed pipeline class has been iterated over. If the number of unique pipelines to search over per batch is one, ensembling will not run. Defaults to False.
- **max_batches** (*int*) – The maximum number of batches of pipelines to search. Parameters max_time, and max_iterations have precedence over stopping the search.
- **problem_configuration** (*dict, None*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the date_index, gap, forecast_horizon, and max_delay variables.
- **train_best_pipeline** (*boolean*) – Whether or not to train the best pipeline before returning it. Defaults to True.
- **pipeline_parameters** (*dict*) – A dict of the parameters used to initialize a pipeline with. Keys should consist of the component names and values should specify parameter values
e.g. pipeline_parameters = { 'Imputer' : { 'numeric_impute_strategy': 'most_frequent' } }
- **custom_hyperparameters** (*dict*) – A dict of the hyperparameter ranges used to iterate over during search. Keys should consist of the component names and values should specify a singular value or skopt.Space.
e.g. custom_hyperparameters = { 'Imputer' : { 'numeric_impute_strategy': Categorical(['most_frequent', 'median']) } }
- **sampler_method** (*str*) – The data sampling component to use in the pipelines if the problem type is classification and the target balance is smaller than the sampler_balanced_ratio. Either 'auto', which will use our preferred sampler for the data, 'Undersampler', 'Oversampler', or None. Defaults to 'auto'.
- **sampler_balanced_ratio** (*float*) – The minority:majority class ratio that we consider balanced, so a 1:4 ratio would be equal to 0.25. If the class balance is larger than this provided value, then we will not add a sampler since the data is then considered balanced. Overrides the *sampler_ratio* of the samplers. Defaults to 0.25.
- **_ensembling_split_size** (*float*) – The amount of the training data we'll set aside for training ensemble metalearners. Only used when ensembling is True. Must be between 0 and 1, exclusive. Defaults to 0.2
- **_pipelines_per_batch** (*int*) – The number of pipelines to train for every batch after the first one. The first batch will train a baseline pipeline + one of each pipeline family allowed in the search.
- **_automl_algorithm** (*str*) – The automl algorithm to use. Currently the two choices are 'iterative' and 'default'. Defaults to *iterative*.

- **engine** (`EngineBase` or `str`) – The engine instance used to evaluate pipelines. Dask or concurrent.futures engines can also be chosen by providing a string from the list [“sequential”, “cf_threaded”, “cf_process”, “dask_threaded”, “dask_process”]. If a parallel engine is selected this way, the maximum amount of parallelism, as determined by the engine, will be used. Defaults to “sequential”.
- **verbose** (`boolean`) – Whether or not to display semi-real-time updates to stdout while search is running. Defaults to False.

Methods

<code>add_to_rankings</code>	Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.
<code>best_pipeline</code>	Returns a trained instance of the best pipeline and parameters found during automl search. If <code>train_best_pipeline</code> is set to False, returns an untrained pipeline instance.
<code>close_engine</code>	Function to explicitly close the engine, client, parallel resources.
<code>describe_pipeline</code>	Describe a pipeline
<code>full_rankings</code>	Returns a pandas.DataFrame with scoring results from all pipelines searched
<code>get_pipeline</code>	Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline
<code>load</code>	Loads AutoML object at file path
<code>plot</code>	
<code>rankings</code>	Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.
<code>results</code>	Class that allows access to a copy of the results from <code>automl_search</code> .
<code>save</code>	Saves AutoML object at file path
<code>score_pipelines</code>	Score a list of pipelines on the given holdout data.
<code>search</code>	Find the best pipeline for the data set.
<code>train_pipelines</code>	Train a list of pipelines on the training data.

add_to_rankings (*self*, *pipeline*)

Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.

Parameters **pipeline** (`PipelineBase`) – pipeline to train and evaluate.

property best_pipeline (*self*)

Returns a trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Returns A trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Return type `PipelineBase`

close_engine (*self*)

Function to explicitly close the engine, client, parallel resources.

describe_pipeline (*self*, *pipeline_id*, *return_dict=False*)

Describe a pipeline

Parameters

- **pipeline_id** (*int*) – pipeline to describe
- **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Description of specified pipeline. Includes information such as type of pipeline components, problem, training time, cross validation, etc.

property full_rankings (*self*)

Returns a pandas.DataFrame with scoring results from all pipelines searched

get_pipeline (*self*, *pipeline_id*)

Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline initialized with the parameters used to train that pipeline during automl search.

Parameters **pipeline_id** (*int*) – pipeline to retrieve

Returns untrained pipeline instance associated with the provided ID

Return type PipelineBase

static load (*file_path*, *pickle_type='cloudpickle'*)

Loads AutoML object at file path

Parameters

- **file_path** (*str*) – location to find file to load
- **{"pickle"}** (*pickle_type*) – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.
- **"cloudpickle"** – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.

Returns AutoSearchBase object

property plot (*self*)

property rankings (*self*)

Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.

property results (*self*)

Class that allows access to a copy of the results from *automl_search*.

Returns: dict containing **pipeline_results**: a dict with results from each pipeline, and **search_order**: a list describing the order the pipelines were searched.

save (*self*, *file_path*, *pickle_type='cloudpickle'*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves AutoML object at file path

Parameters

- **file_path** (*str*) – location to save file
- **{"pickle"}** (*pickle_type*) – the pickling library to use.
- **"cloudpickle"** – the pickling library to use.
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score_pipelines (*self*, *pipelines*, *X_holdout*, *y_holdout*, *objectives*)

Score a list of pipelines on the given holdout data.

Parameters

- **pipelines** (*list* (*PipelineBase*)) – List of pipelines to train.
- **X_holdout** (*pd.DataFrame*) – Holdout features.
- **y_holdout** (*pd.Series*) – Holdout targets for scoring.
- **objectives** (*list* (*str*), *list* (*ObjectiveBase*)) – Objectives used for scoring.

Returns Dictionary keyed by pipeline name that maps to a dictionary of scores. Note that the any pipelines that error out during scoring will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, Dict[str, float]]

search (*self*, *show_iteration_plot=True*)

Find the best pipeline for the data set.

Parameters

- **feature_types** (*list*, *optional*) – list of feature types, either numerical or categorical. Categorical features will automatically be encoded
- **show_iteration_plot** (*boolean*, *True*) – Shows an iteration vs. score plot in Jupyter notebook. Disabled by default in non-Jupyter environments.

train_pipelines (*self*, *pipelines*)

Train a list of pipelines on the training data.

This can be helpful for training pipelines once the search is complete.

Parameters **pipelines** (*list* (*PipelineBase*)) – List of pipelines to train.

Returns Dictionary keyed by pipeline name that maps to the fitted pipeline. Note that the any pipelines that error out during training will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, PipelineBase]

class evalml.automl.EngineBase

Helper class that provides a standard way to create an ABC using inheritance.

Methods

setup_job_log

submit_evaluation_job

Submit job for pipeline evaluation during AutoMLSearch.

submit_scoring_job

Submit job for pipeline scoring.

submit_training_job

Submit job for pipeline training.

static **setup_job_log** ()

abstract **submit_evaluation_job** (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline evaluation during AutoMLSearch.

abstract **submit_scoring_job** (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*)

Submit job for pipeline scoring.

abstract submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline training.

evalml.automl.get_default_primary_search_objective (*problem_type*)

Get the default primary search objective for a problem type.

Parameters *problem_type* (*str* or *ProblemType*) – problem type of interest.

Returns primary objective instance for the problem type.

Return type ObjectiveBase

evalml.automl.make_data_splitter (*X*, *y*, *problem_type*, *problem_configuration=None*,
n_splits=3, *shuffle=True*, *random_seed=0*)

Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.

Parameters

- **X** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].
- **problem_type** (*ProblemType*) – The type of machine learning problem.
- **problem_configuration** (*dict*, *None*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the *date_index*, *gap*, and *max_delay* variables. Defaults to *None*.
- **n_splits** (*int*, *None*) – The number of CV splits, if applicable. Defaults to 3.
- **shuffle** (*bool*) – Whether or not to shuffle the data before splitting, if applicable. Defaults to *True*.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Data splitting method.

Return type sklearn.model_selection.BaseCrossValidator

evalml.automl.search (*X_train=None*, *y_train=None*, *problem_type=None*, *objective='auto'*,
mode='fast', *max_time=None*, *patience=None*, *tolerance=None*, *problem_configuration=None*)

Given data and configuration, run an automl search.

This method will run EvalML's default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you'd like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.
- **mode** (*str*) – mode for DefaultAlgorithm. There are two modes: fast and long, where fast is a subset of long. Please look at DefaultAlgorithm for more details.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in seconds. For strings, time can be specified as seconds, minutes, or hours.
- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If None, early stopping is disabled. Defaults to None.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if patience is not None. Defaults to None.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, and `max_delay` variables.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

```
evalml.automl.search_iterative(X_train=None, y_train=None, problem_type=None, objective='auto', problem_configuration=None, **kwargs)
```

Given data and configuration, run an automl search.

This method will run EvalML’s default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you’d like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.
- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to ‘auto’, chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.

- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example,
- **time series problems** (*in*) –
- **should be passed in for the date_index** (*values*) –
- **gap** –
- **forecast_horizon** –
- **max_delay variables.** (*and*) –

Other keyword arguments which are provided will be passed to AutoMLSearch.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

class evalml.automl.SequentialEngine

The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Methods

close

setup_job_log

<i>submit_evaluation_job</i>	Submit job for pipeline evaluation during AutoMLSearch.
------------------------------	---

<i>submit_scoring_job</i>	Submit job for pipeline scoring.
---------------------------	----------------------------------

<i>submit_training_job</i>	Submit job for pipeline training.
----------------------------	-----------------------------------

close (*self*)

static setup_job_log ()

submit_evaluation_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline evaluation during AutoMLSearch.

submit_scoring_job (*self*, *automl_config*, *pipeline*, *X*, *y*, *objectives*)

Submit job for pipeline scoring.

submit_training_job (*self*, *automl_config*, *pipeline*, *X*, *y*)

Submit job for pipeline training.

evalml.automl.tune_binary_threshold (*pipeline*, *objective*, *problem_type*, *X_threshold_tuning*, *y_threshold_tuning*)

Tunes the threshold of a binary pipeline to the X and y thresholding data

Parameters

- **pipeline** (*Pipeline*) – Pipeline instance to threshold.
- **objective** (*ObjectiveBase*) – The objective we want to tune with. If not tuneable and best_pipeline is True, will use F1.
- **problem_type** (*ProblemType*) – The problem type of the pipeline.
- **X_threshold_tuning** (*pd.DataFrame*) – Features to tune pipeline to.
- **y_threshold_tuning** (*pd.Series*) – Target data to tune pipeline to.

Data Checks

Submodules

`class_imbalance_data_check`

Module Contents

Classes Summary

<code>ClassImbalanceDataCheck</code>	Check if any of the target labels are imbalanced, or if the number of values for each target are below 2 times the number of CV folds. Use for classification problems.
--------------------------------------	---

Contents

class `evalml.data_checks.class_imbalance_data_check.ClassImbalanceDataCheck` (*threshold=0.1, min_samples=100, num_cv_folds=3*)

Check if any of the target labels are imbalanced, or if the number of values for each target are below 2 times the number of CV folds. Use for classification problems.

Parameters

- **threshold** (*float*) – The minimum threshold allowed for class imbalance before a warning is raised. This threshold is calculated by comparing the number of samples in each class to the sum of samples in that class and the majority class. For example, a multi-class case with [900, 900, 100] samples per classes 0, 1, and 2, respectively, would have a 0.10 threshold for class 2 ($100 / (900 + 100)$). Defaults to 0.10.
- **min_samples** (*int*) – The minimum number of samples per accepted class. If the minority class is both below the threshold and min_samples, then we consider this severely imbalanced. Must be greater than 0. Defaults to 100.
- **num_cv_folds** (*int*) – The number of cross-validation folds. Must be positive. Choose 0 to ignore this warning. Defaults to 3.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Checks if any target labels are imbalanced beyond a threshold for binary and multiclass problems

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y*)

Checks if any target labels are imbalanced beyond a threshold for binary and multiclass problems
Ignores NaN values in target labels if they appear.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features. Ignored.

- `y` (`pd.Series`, `np.ndarray`) – Target labels to check for imbalanced data.

Returns

Dictionary with `DataCheckWarnings` if imbalance in classes is less than the threshold, and `DataCheckErrors` if the number of values for each target is below `2 * num_cv_folds`.

Return type dict

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame()
>>> y = pd.Series([0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
>>> target_check = ClassImbalanceDataCheck(threshold=0.10)
>>> assert target_check.validate(X, y) == {"errors": [{"message": "The number_
↳ of instances of these targets is less than 2 * the number of cross folds =_
↳ 6 instances: [0]",
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "error",
...                                     "code": "CLASS_
↳ IMBALANCE_BELOW_FOLDS",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     "warnings": [{"message": "The_
↳ following labels fall below 10% of the target: [0]",
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "warning",
...                                     "code": "CLASS_
↳ IMBALANCE_BELOW_THRESHOLD",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     {"message": "The_
↳ following labels in the target have severe class imbalance because they_
↳ fall under 10% of the target and have less than 100 samples: [0]",
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "warning",
...                                     "code": "CLASS_
↳ IMBALANCE_SEVERE",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     "actions": []}]}
```

`data_check`

Module Contents

Classes Summary

<i>DataCheck</i>	Base class for all data checks. Data checks are a set of heuristics used to determine if there are problems with input data.
------------------	--

Contents

class evalml.data_checks.data_check.DataCheck

Base class for all data checks. Data checks are a set of heuristics used to determine if there are problems with input data.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Inspects and validates the input data, runs any necessary calculations or algorithms, and returns a list of warnings and errors if applicable.

name (*cls*)

Returns a name describing the data check.

abstract validate (*self*, *X*, *y=None*)

Inspects and validates the input data, runs any necessary calculations or algorithms, and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target data of length [n_samples]

Returns Dictionary of DataCheckError and DataCheckWarning messages

Return type dict (DataCheckMessage)

data_check_action

Module Contents

Classes Summary

<i>DataCheckAction</i>	A recommended action returned by a DataCheck.
------------------------	---

Contents

class evalml.data_checks.data_check_action.**DataCheckAction** (*action_code*, *meta-data=None*)

A recommended action returned by a DataCheck.

Parameters

- **action_code** (*DataCheckActionCode*) – Action code associated with the action.
- **metadata** (*dict*, *optional*) – Additional useful information associated with the action. Defaults to None.

Methods

to_dict

to_dict (*self*)

data_check_action_code

Module Contents

Classes Summary

<i>DataCheckActionCode</i>	Enum for data check action code.
----------------------------	----------------------------------

Contents

class evalml.data_checks.data_check_action_code.**DataCheckActionCode**
Enum for data check action code.

Attributes

DROP_COL	Action code for dropping a column.
DROP_ROWS	Action code for dropping rows.
IM- PUTE_COL	Action code for imputing a column.
TRANS- FORM_TARGET	Action code for transforming the target data.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)
The name of the Enum member.

value (*self*)

The value of the Enum member.

data_check_message

Module Contents

Classes Summary

<i>DataCheckError</i>	DataCheckMessage subclass for errors returned by data checks.
<i>DataCheckMessage</i>	Base class for a message returned by a DataCheck, tagged by name.
<i>DataCheckWarning</i>	DataCheckMessage subclass for warnings returned by data checks.

Contents

class evalml.data_checks.data_check_message.**DataCheckError** (*message*,
data_check_name,
message_code=None,
details=None)

DataCheckMessage subclass for errors returned by data checks.

Attributes

mes- sage_type	DataCheckMessageType.ERROR
---------------------------	----------------------------

Methods

to_dict

to_dict (*self*)

class evalml.data_checks.data_check_message.**DataCheckMessage** (*message*,
data_check_name,
*mes-
sage_code=None*,
details=None)

Base class for a message returned by a DataCheck, tagged by name.

Parameters

- **message** (*str*) – Message string
- **data_check_name** (*str*) – Name of data check
- **message_code** (*DataCheckMessageCode*) – Message code associated with message. Defaults to None.
- **details** (*dict*) – Additional useful information associated with the message. Defaults to None.

Attributes

mes- sage_type	None
-------------------	------

Methods

to_dict

to_dict (*self*)

```
class evalml.data_checks.data_check_message.DataCheckWarning (message,
                                                                data_check_name,
                                                                mes-
                                                                sage_code=None,
                                                                details=None)
```

DataCheckMessage subclass for warnings returned by data checks.

Attributes

mes- sage_type	DataCheckMessageType.WARNING
-------------------	------------------------------

Methods

to_dict

to_dict (*self*)

data_check_message_code

Module Contents

Classes Summary

<i>DataCheckMessageCode</i>	Enum for data check message code.
-----------------------------	-----------------------------------

Contents

class evalml.data_checks.data_check_message_code.**DataCheckMessageCode**
Enum for data check message code.

Attributes

CLASS_IMBALANCE_BELOW_FOLDS	number of values for each target is below 2 * number of CV folds.
------------------------------------	---

continues on next page

Table 95 – continued from previous page

CLASS_IMBALANCE_BELOW_THRESHOLD	Message code for when balance in classes is less than the threshold.
CLASS_IMBALANCE_SEVERE	Message code for when balance in classes is less than the threshold and minimum class is less than minimum number of accepted samples.
DATE-TIME_HAS_NAN	Message code for when input datetime columns contain NaN values.
DATE-TIME_HAS_UNEVEN_INTERVALS	Message code for when the datetime values have uneven intervals.
DATE-TIME_INFORMATION_NOT_FOUND	Message code for when datetime information can not be found or is in an unaccepted format.
DATE-TIME_IS_NOT_MONOTONIC	Message code for when the datetime values are not monotonically increasing.
HAS_ID_COLUMNS	Message code for data that has ID columns.
HAS_OUTLIERS	Message code for when outliers are detected.
HIGH_VARIANCE	Message code for when high variance is detected for cross-validation.
HIGHLY_NULL_COLUMNS	Message code for highly null columns.
HIGHLY_NULL_ROWS	Message code for highly null rows.
IS_MULTICOLLINEAR	Message code for when data is potentially multicollinear.
MIS-MATCHED_INDICES	Message code for when input target and features have mismatched indices.
MIS-MATCHED_INDICES_ORDER	Message code for when input target and features have mismatched indices order. The two have the same index values, but shuffled.
MIS-MATCHED_LENGTHS	Message code for when input target and features have different lengths.
NATURAL_LANGUAGE_HAS_NAN	Message code for when input natural language columns contain NaN values.
NO_VARIANCE	Message code for when data has no variance (1 unique value).
NO_VARIANCE_WITH_NULL	Message code for when data has one unique value and NaN values.
NOT_UNIQUE_ENOUGH	Message code for when data does not possess enough unique values.
TARGET_BINARY_NOT_TWO_UNIQUE_VALUES	Message code for target data for a binary classification problem that does not have two unique values.
TARGET_HAS_NULL	Message code for target data that has null values.
TARGET_INCOMPATIBLE_OBJECTIVE	Message code for target data that has incompatible values for the specified objective
TARGET_IS_EMPTY_OR_FULLY_NULL	Message code for target data that is empty or has all null values.
TARGET_IS_NONE	Message code for when target is None.
TARGET_LEAKAGE	Message code for when target leakage is detected.
TARGET_LOGNORMAL_DISTRIBUTION	Message code for target data with a lognormal distribution.
TARGET_MULTICLASS_HIGH_UNIQUE_CLASSES	Message code for target data for a multi classification problem that has an abnormally large number of target values.
TARGET_MULTICLASS_NOT_ENOUGH_CLASSES	Message code for target data for a multi classification problem that does not have more than a minimum number of classes.
TARGET_MULTICLASS_NOT_TWO_EXAMPLES_PER_CLASS	Message code for target data for a multi classification problem that does not have two examples per class.
TARGET_UNSUPPORTED_PROBLEM_TYPE	Message code for target data that is being checked against an unsupported problem type.

continues on next page

Table 95 – continued from previous page

TAR-GET_UNSUPPORTED_TYPE	Message code for target data that is of an unsupported type.
TOO_SPARSE	Message code for when multiclass data has values that are too sparsely populated.
TOO_UNIQUE	Message code for when data possesses too many unique values.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

data_check_message_type**Module Contents****Classes Summary**

<i>DataCheckMessageType</i>	Enum for type of data check message: WARNING or ERROR.
-----------------------------	--

Contents**class** evalml.data_checks.data_check_message_type.**DataCheckMessageType**

Enum for type of data check message: WARNING or ERROR.

Attributes

ERROR	Error message returned by a data check.
WARNING	Warning message returned by a data check.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

data_checks

Module Contents

Classes Summary

<i>DataChecks</i>	A collection of data checks.
-------------------	------------------------------

Contents

class evalml.data_checks.data_checks.**DataChecks** (*data_checks=None*,
data_check_params=None)

A collection of data checks.

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self*, *X*, *y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

datetime_format_data_check

Module Contents

Classes Summary

<i>DateTimeFormatDataCheck</i>	Checks if the datetime column has equally spaced intervals and is monotonically increasing or decreasing in order
--------------------------------	---

Contents

class evalml.data_checks.datetime_format_data_check.**DateTimeFormatDataCheck** (*datetime_column='in*

Checks if the datetime column has equally spaced intervals and is monotonically increasing or decreasing in order to be supported by time series estimators.

Parameters **datetime_column** (*str, int*) – The name of the datetime column. If the date-time values are in the index, then pass “index”.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if the target data has equal intervals and is sorted.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y*)

Checks if the target data has equal intervals and is sorted.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Target data.

Returns List with DataCheckErrors if unequal intervals are found in the datetime column.

Return type dict (DataCheckError)

Example

```
>>> from pandas as pd
>>> X = pd.DataFrame(pd.date_range("January 1, 2021", periods=8), columns=[
↳ "dates"])
>>> y = pd.Series([1, 2, 4, 2, 1, 2, 3, 1])
>>> X.iloc[7] = "January 9, 2021"
>>> datetime_format_check = DateTimeFormatDataCheck()
>>> assert datetime_format_check.validate(X, y) == {
...     "errors": [{"message": "No frequency could be detected in dates,
↳ possibly due to uneven intervals.",
...                 "data_check_name": "EqualIntervalDataCheck",
...                 "level": "error",
...                 "code": "DATETIME_HAS_UNEVEN_INTERVALS",
...                 "details": {}}],
...     "warnings": [],
...     "actions": []}
```

datetime_nan_data_check

Module Contents

Classes Summary

<i>DateTimeNaNDataCheck</i>	Checks each column in the input for datetime features and will issue an error if NaN values are present.
-----------------------------	--

Attributes Summary

<i>error_contains_nan</i>

Contents

class evalml.data_checks.datetime_nan_data_check.**DateTimeNaNDataCheck**
Checks each column in the input for datetime features and will issue an error if NaN values are present.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if any datetime columns contain NaN values.

name (*cls*)
Returns a name describing the data check.

validate (*self*, *X*, *y=None*)
Checks if any datetime columns contain NaN values.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features.
- **y** (*pd.Series*, *np.ndarray*) – Ignored. Defaults to None.

Returns dict with a DataCheckError if NaN values are present in datetime columns.

Return type dict

Example

```
>>> import pandas as pd
>>> import woodwork as ww
>>> import numpy as np
>>> dates = np.arange(np.datetime64('2017-01-01'), np.datetime64('2017-01-08
↪'))
>>> dates[0] = np.datetime64('NaT')
>>> df = pd.DataFrame(dates, columns=['index'])
>>> df.ww.init()
>>> dt_nan_check = DateTimeNaNDataCheck()
```

(continues on next page)

(continued from previous page)

```

>>> assert dt_nan_check.validate(df) == {"warnings": [],
...                                     "actions": [],
...                                     "errors": [
↪ [DataCheckError(message='Input datetime column(s) (index) contains NaN',
↪ values. Please impute NaN values or drop these rows or columns.',
...                                     data_
↪ check_name=DateTimeNaNDataCheck.name,
...                                     ]
↪ message_code=DataCheckMessageCode.DATETIME_HAS_NAN,
...                                     ]
↪ details={"columns": 'index'}).to_dict()]

```

```
evalml.data_checks.datetime_nan_data_check.error_contains_nan = Input datetime column(s) (
```

default_data_checks

Module Contents

Classes Summary

DefaultDataChecks

A collection of basic data checks that is used by AutoML by default.

Contents

class evalml.data_checks.default_data_checks.**DefaultDataChecks** (*problem_type*,
objective,
n_splits=3,
date-
time_column=None)

A collection of basic data checks that is used by AutoML by default. Includes:

- *HighlyNullDataCheck*
- *HighlyNullRowsDataCheck*
- *IDColumnsDataCheck*
- *TargetLeakageDataCheck*
- *InvalidTargetDataCheck*
- *NoVarianceDataCheck*
- *ClassImbalanceDataCheck* (for classification problem types)
- *DateTimeNaNDataCheck*
- *NaturalLanguageNaNDataCheck*
- *TargetDistributionDataCheck* (for regression problem types)
- *DateTimeFormatDataCheck* (for time series problem types)

Parameters

- **problem_type** (*str*) – The problem type that is being validated. Can be regression, binary, or multiclass.
- **objective** (*str* or *ObjectiveBase*) – Name or instance of the objective class.
- **n_splits** (*int*) – The number of splits as determined by the data splitter being used. Defaults to 3.
- **datetime_column** (*str*) – The name of the column containing datetime information to be used for time series problems.
- **to "index" indicating that the datetime information is in the index of X or y. (Default)** –

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self*, *X*, *y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

highly_null_data_check

Module Contents

Classes Summary

<i>HighlyNullDataCheck</i>	Checks if there are any highly-null columns and rows in the input.
----------------------------	--

Contents

class evalml.data_checks.highly_null_data_check.**HighlyNullDataCheck** (*pct_null_col_threshold=0.95*, *pct_null_row_threshold=0.95*)

Checks if there are any highly-null columns and rows in the input.

Parameters

- **pct_null_col_threshold** (*float*) – If the percentage of NaN values in an input feature exceeds this amount, that column will be considered highly-null. Defaults to 0.95.
- **pct_null_row_threshold** (*float*) – If the percentage of NaN values in an input row exceeds this amount, that row will be considered highly-null. Defaults to 0.95.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Checks if there are any highly-null columns or rows in the input.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Checks if there are any highly-null columns or rows in the input.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Ignored.

Returns dict with a DataCheckWarning if there are any highly-null columns or rows.

Return type dict

Example

```
>>> import pandas as pd
>>> class SeriesWrap():
...     def __init__(self, series):
...         self.series = series
...
...     def __eq__(self, series_2):
...         return all(self.series.eq(series_2.series))
...
>>> df = pd.DataFrame({
...     'lots_of_null': [None, None, None, None, 5],
...     'no_null': [1, 2, 3, 4, 5]
... })
>>> null_check = HighlyNullDataCheck(pct_null_col_threshold=0.50, pct_null_
↪row_threshold=0.50)
>>> validation_results = null_check.validate(df)
>>> validation_results['warnings'][0]['details']['pct_null_cols'] =_
↪SeriesWrap(validation_results['warnings'][0]['details']['pct_null_cols'])
>>> highly_null_rows = SeriesWrap(pd.Series([0.5, 0.5, 0.5, 0.5]))
>>> assert validation_results == {
...     "errors": [],
...     "warnings": [{
...         "message": "4 out of 5 rows are more than 50.0% null",
...         "data_check_name": "HighlyNullDataCheck",
...         "level": "warning",
...         "code": "HIGHLY_NULL_ROWS",
...         "details": {"pct_null_cols": highly_null_rows}},
...         {"message": "Column 'lots_of_null' is 50.0% or more null
↪",
...         "data_check_name": "HighlyNullDataCheck",
...         "level": "warning",
...         "code": "HIGHLY_NULL_COLS",
...         "details": {"column": "lots_of_null", "pct_null_rows":_
↪0.8}}}],
...     "actions": [{"code": "DROP_ROWS", "metadata": {"rows": [0, 1, 2, 3]}},
...         {"code": "DROP_COL", "metadata": {"column": "lots_of_null"}}
...     ]
↪}
```

(continues on next page)

(continued from previous page)

id_columns_data_check

Module Contents

Classes Summary

<code>IDColumnsDataCheck</code>	Check if any of the features are likely to be ID columns.
---------------------------------	---

Contents

class `evalml.data_checks.id_columns_data_check.IDColumnsDataCheck` (*id_threshold=1.0*)
Check if any of the features are likely to be ID columns.

Parameters `id_threshold` (*float*) – The probability threshold to be considered an ID column.
Defaults to 1.0.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Check if any of the features are likely to be ID columns. Currently performs these simple checks:

name (*cls*)
Returns a name describing the data check.

validate (*self, X, y=None*)
Check if any of the features are likely to be ID columns. Currently performs these simple checks:

- column name is “id”
- column name ends in “_id”
- column contains all unique values (and is categorical / integer type)

Parameters `X` (*pd.DataFrame, np.ndarray*) – The input features to check

Returns A dictionary of features with column name or index and their probability of being ID columns

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'df_id': [0, 1, 2, 3, 4],
...     'x': [10, 42, 31, 51, 61],
...     'y': [42, 54, 12, 64, 12]
... })
>>> id_col_check = IDColumnsDataCheck()
>>> assert id_col_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Column 'df_id' is 100.0% or more likely to_
↳ be an ID column",
...                     "data_check_name": "IDColumnsDataCheck",
...                     "level": "warning",
...                     "code": "HAS_ID_COLUMN",
...                     "details": {"column": "df_id"}}],
...     "actions": [{"code": "DROP_COL",
...                   "metadata": {"column": "df_id"}}]}
```

invalid_targets_data_check

Module Contents

Classes Summary

InvalidTargetDataCheck

Checks if the target data contains missing or invalid values.

Contents

class evalml.data_checks.invalid_targets_data_check.**InvalidTargetDataCheck** (*problem_type*, *ob-*
jec-
tive, *n_unique=100*)

Checks if the target data contains missing or invalid values.

Parameters

- **problem_type** (*str* or *ProblemTypes*) – The specific problem type to data check for. e.g. ‘binary’, ‘multiclass’, ‘regression’, ‘time series regression’
- **objective** (*str* or *ObjectiveBase*) – Name or instance of the objective class.
- **n_unique** (*int*) – Number of unique target values to store when problem type is binary and target incorrectly has more than 2 unique values. Non-negative integer. If None, stores all unique values. Defaults to 100.

Attributes

multi-class_continuous_threshold	0.05
----------------------------------	------

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Checks if the target data contains missing or invalid values.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Checks if the target data contains missing or invalid values.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features. Ignored.
- **y** (*pd.Series*, *np.ndarray*) – Target data to check for invalid values.

Returns List with `DataCheckErrors` if any invalid values are found in the target data.

Return type dict (`DataCheckError`)

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame({"col": [1, 2, 3, 1]})
>>> y = pd.Series([0, 1, None, None])
>>> target_check = InvalidTargetDataCheck('binary', 'Log Loss Binary')
>>> assert target_check.validate(X, y) == {
...     "errors": [{"message": "2 row(s) (50.0%) of target values are null",
...                     "data_check_name": "InvalidTargetDataCheck",
...                     "level": "error",
...                     "code": "TARGET_HAS_NULL",
...                     "details": {"num_null_rows": 2, "pct_null_rows": 50}}],
...     "warnings": [],
...     "actions": [{"code": 'IMPUTE_COL', 'metadata': {'column': None,
...     ↪ 'impute_strategy': 'most_frequent', 'is_target': True}}]}
```

multicollinearity_data_check

Module Contents

Classes Summary

<code>MulticollinearityDataCheck</code>	Check if any set features are likely to be multicollinear.
---	--

Contents

class evalml.data_checks.multicollinearity_data_check.**MulticollinearityDataCheck** (*threshold=0.9*)
Check if any set features are likely to be multicollinear.

Parameters **threshold** (*float*) – The threshold to be considered. Defaults to 0.9.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if any set of features are likely to be multicollinear.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Check if any set of features are likely to be multicollinear.

Parameters

- **X** (*pd.DataFrame*) – The input features to check.
- **y** (*pd.Series*) – The target. Ignored.

Returns dict with a DataCheckWarning if there are any potentially multicollinear columns.

Return type dict

Example

```
>>> import pandas as pd
>>> col = pd.Series([1, 0, 2, 3, 4])
>>> X = pd.DataFrame({"col_1": col, "col_2": col * 3})
>>> y = pd.Series([1, 0, 0, 1, 0])
>>> multicollinearity_check = MulticollinearityDataCheck(threshold=0.8)
>>> assert multicollinearity_check.validate(X, y) == {
...     "errors": [],
...     "warnings": [{"message": "Columns are likely to be correlated: [('col_
↪ 1', 'col_2')]",
...                     "data_check_name": "MulticollinearityDataCheck",
...                     "level": "warning",
...                     "code": "IS_MULTICOLLINEAR",
...                     "details": {'columns': [('col_1', 'col_2')]}},
...     "actions": []}
```

natural_language_nan_data_check

Module Contents

Classes Summary

NaturalLanguageNaNDataCheck

Checks each column in the input for natural language features and will issue an error if NaN values are present.

Attributes Summary

error_contains_nan

Contents

`evalml.data_checks.natural_language_nan_data_check.error_contains_nan = Input natural language`

class `evalml.data_checks.natural_language_nan_data_check.NaturalLanguageNaNDataCheck`

Checks each column in the input for natural language features and will issue an error if NaN values are present.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if any natural language columns contain NaN values.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Checks if any natural language columns contain NaN values.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Ignored. Defaults to None.

Returns dict with a DataCheckError if NaN values are present in natural language columns.

Return type dict

Example

```
>>> import pandas as pd
>>> import woodwork as ww
>>> import numpy as np
>>> data = pd.DataFrame()
>>> data['A'] = [None, "string_that_is_long_enough_for_natural_language"]
>>> data['B'] = ['string_that_is_long_enough_for_natural_language', 'string_
↳ that_is_long_enough_for_natural_language']
>>> data['C'] = np.random.randint(0, 3, size=len(data))
>>> data.ww.init(logical_types={'A': 'NaturalLanguage', 'B': 'NaturalLanguage
↳ '})
>>> nl_nan_check = NaturalLanguageNaNDataCheck()
>>> assert nl_nan_check.validate(data) == {
...     "warnings": [],
```

(continues on next page)

(continued from previous page)

```

...         "actions": [],
...         "errors": [DataCheckError(message='Input natural language_
↪column(s) (A) contains NaN values. Please impute NaN values or drop these_
↪rows or columns.',
...                                     data_check_name=NaturalLanguageNaNDataCheck.name,
...                                     message_code=DataCheckMessageCode.NATURAL_LANGUAGE_
↪HAS_NAN,
...                                     details={"columns": 'A'}).to_dict()]
...     }

```

no_variance_data_check

Module Contents

Classes Summary

<i>NoVarianceDataCheck</i>	Check if the target or any of the features have no variance.
----------------------------	--

Contents

class evalml.data_checks.no_variance_data_check.**NoVarianceDataCheck** (*count_nan_as_value=False*)
 Check if the target or any of the features have no variance.

Parameters *count_nan_as_value* (*bool*) – If True, missing values will be counted as their own unique value. Additionally, if true, will return a DataCheckWarning instead of an error if the feature has mostly missing data and only one unique value. Defaults to False.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if the target or any of the features have no variance (1 unique value).

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y*)

Check if the target or any of the features have no variance (1 unique value).

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input features.
- **y** (*pd.Series, np.ndarray*) – The target data.

Returns dict of warnings/errors corresponding to features or target with no variance.

Return type dict

outliers_data_check

Module Contents

Classes Summary

<i>OutliersDataCheck</i>	Checks if there are any outliers in input data by using IQR to determine score anomalies. Columns with score anomalies are considered to contain outliers.
--------------------------	--

Contents

class evalml.data_checks.outliers_data_check.OutliersDataCheck

Checks if there are any outliers in input data by using IQR to determine score anomalies. Columns with score anomalies are considered to contain outliers.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if there are any outliers in a dataframe by using IQR to determine column anomalies. Column with anomalies are considered to contain outliers.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Checks if there are any outliers in a dataframe by using IQR to determine column anomalies. Column with anomalies are considered to contain outliers.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features
- **y** (*pd.Series, np.ndarray*) – Ignored.

Returns A dictionary with warnings if any columns have outliers.

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'x': [1, 2, 3, 4, 5],
...     'y': [6, 7, 8, 9, 10],
...     'z': [-1, -2, -3, -1201, -4]
... })
>>> outliers_check = OutliersDataCheck()
>>> assert outliers_check.validate(df) == {
...     "warnings": [{"message": "Column(s) 'z' are likely to have outlier",
... data.",
...     "data_check_name": "OutliersDataCheck",
...     "level": "warning",
... }
```

(continues on next page)

(continued from previous page)

```

...         "code": "HAS_OUTLIERS",
...         "details": {"columns": ["z"]}},
...     "errors": [],
...     "actions": []}

```

sparsity_data_check

Module Contents

Classes Summary

<i>SparsityDataCheck</i>	Checks if there are any columns with sparsely populated values in the input.
--------------------------	--

Attributes Summary

<i>warning_too_unique</i>	
---------------------------	--

Contents

class evalml.data_checks.sparsity_data_check.**SparsityDataCheck** (*problem_type*, *threshold*, *unique_count_threshold=10*)

Checks if there are any columns with sparsely populated values in the input.

Parameters

- **problem_type** (*str* or *ProblemTypes*) – The specific problem type to data check for. ‘multiclass’ or ‘time series multiclass’ is the only accepted problem type.
- **threshold** (*float*) – The threshold value, or percentage of each column’s unique values, below which, a column exhibits sparsity. Should be between 0 and 1.
- **unique_count_threshold** (*int*) – The minimum number of times a unique value has to be present in a column to not be considered “sparse.” Defaults to 10.

Methods

<i>name</i>	Returns a name describing the data check.
<i>sparsity_score</i>	This function calculates a sparsity score for the given value counts by calculating the percentage of
<i>validate</i>	Calculates what percentage of each column’s unique values exceed the count threshold and compare

name (*cls*)

Returns a name describing the data check.

static sparsity_score (*col*, *count_threshold=10*)

This function calculates a sparsity score for the given value counts by calculating the percentage of unique values that exceed the `count_threshold`.

Parameters

- `col` (*pd.Series*) – Feature values.
- `count_threshold` (*int*) – The number of instances below which a value is considered sparse. Default is 10.

Returns Sparsity score, or the percentage of the unique values that exceed `count_threshold`.

Return type (float)

validate (*self, X, y=None*)

Calculates what percentage of each column's unique values exceed the count threshold and compare that percentage to the sparsity threshold stored in the class instance.

Parameters

- `X` (*pd.DataFrame, np.ndarray*) – Features.
- `y` (*pd.Series, np.ndarray*) – Ignored.

Returns dict with a `DataCheckWarning` if there are any sparse columns.

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'sparse': [float(x) for x in range(100)],
...     'not_sparse': [float(1) for x in range(100)]
... })
>>> sparsity_check = SparsityDataCheck(problem_type="multiclass", threshold=0.
↪5, unique_count_threshold=10)
>>> assert sparsity_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Input columns (sparse) for multiclass_
↪problem type are too sparse.",
...                     "data_check_name": "SparsityDataCheck",
...                     "level": "warning",
...                     "code": "TOO_SPARSE",
...                     "details": {"column": "sparse", 'sparsity_score': 0.0}}
↪],
...     "actions": [{"code": "DROP_COL",
...                   "metadata": {"column": "sparse"}}}]}
```

`evalml.data_checks.sparsity_data_check.warning_too_unique = Input columns ({} for {} prob`

target_distribution_data_check

Module Contents

Classes Summary

<i>TargetDistributionDataCheck</i>	Checks if the target data contains certain distributions that may need to be transformed prior training to
------------------------------------	--

Contents

class evalml.data_checks.target_distribution_data_check.**TargetDistributionDataCheck**
Checks if the target data contains certain distributions that may need to be transformed prior training to improve model performance.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if the target data has a certain distribution.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Checks if the target data has a certain distribution.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features. Ignored.
- **y** (*pd.Series*, *np.ndarray*) – Target data to check for underlying distributions.

Returns List with DataCheckErrors if certain distributions are found in the target data.

Return type dict (DataCheckError)

Example

```
>>> from scipy.stats import lognorm
>>> X = None
>>> y = [0.946, 0.972, 1.154, 0.954, 0.969, 1.222, 1.038, 0.999, 0.973, 0.897]
>>> target_check = TargetDistributionDataCheck()
>>> assert target_check.validate(X, y) == {
...     "errors": [],
...     "warnings": [{"message": "Target may have a lognormal distribution.",
...                          "data_check_name": "TargetDistributionDataCheck",
...                          "level": "warning",
...                          "code": "TARGET_LOGNORMAL_DISTRIBUTION",
...                          "details": {"shapiro-statistic/pvalue": '0.84/0.045'}}],
...     "actions": [{"code": 'TRANSFORM_TARGET', 'metadata': {'column': None,
...     ↪ 'transformation_strategy': 'lognormal', 'is_target': True}}]}
```

target_leakage_data_check

Module Contents

Classes Summary

<i>TargetLeakageDataCheck</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.
-------------------------------	--

Contents

class evalml.data_checks.target_leakage_data_check.**TargetLeakageDataCheck** (*pct_corr_threshold=0.95*, *method='mutual'*)

Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

If *method='mutual'*, this data check uses mutual information and supports all target and feature types. Otherwise, if *method='pearson'*, it uses Pearson correlation and only supports binary with numeric and boolean dtypes. Pearson correlation returns a value in $[-1, 1]$, while mutual information returns a value in $[0, 1]$.

Parameters

- **pct_corr_threshold** (*float*) – The correlation threshold to be considered leakage. Defaults to 0.95.
- **method** (*string*) – The method to determine correlation. Use ‘mutual’ for mutual information, otherwise ‘pearson’ for Pearson correlation. Defaults to ‘mutual’.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y*)

Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

If *method='mutual'*, supports all target and feature types. Otherwise, if *method='pearson'* only supports binary with numeric and boolean dtypes. Pearson correlation returns a value in $[-1, 1]$, while mutual information returns a value in $[0, 1]$.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input features to check
- **y** (*pd.Series, np.ndarray*) – The target data

Returns dict with a DataCheckWarning if target leakage is detected.

Return type dict (DataCheckWarning)

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame({
...     'leak': [10, 42, 31, 51, 61],
...     'x': [42, 54, 12, 64, 12],
...     'y': [13, 5, 13, 74, 24],
... })
>>> y = pd.Series([10, 42, 31, 51, 40])
>>> target_leakage_check = TargetLeakageDataCheck(pct_corr_threshold=0.95)
>>> assert target_leakage_check.validate(X, y) == {
...     "warnings": [{"message": "Column 'leak' is 95.0% or more correlated_
↳ with the target",
...                     "data_check_name": "TargetLeakageDataCheck",
...                     "level": "warning",
...                     "code": "TARGET_LEAKAGE",
...                     "details": {"column": "leak"}}],
...     "errors": [],
...     "actions": [{"code": "DROP_COL",
...                   "metadata": {"column": "leak"}}]}
```

uniqueness_data_check

Module Contents

Classes Summary

<i>UniquenessDataCheck</i>	Checks if there are any columns in the input that are either too unique for classification problems
----------------------------	---

Attributes Summary

<i>warning_not_unique_enough</i>
<i>warning_too_unique</i>

Contents

class evalml.data_checks.uniqueness_data_check.**UniquenessDataCheck** (*problem_type*, *thresh-*
old=0.5)

Checks if there are any columns in the input that are either too unique for classification problems or not unique enough for regression problems.

Parameters

- **problem_type** (*str* or *ProblemTypes*) – The specific problem type to data check for. e.g. ‘binary’, ‘multiclass’, ‘regression’, ‘time series regression’

- **threshold** (*float*) – The threshold to set as an upper bound on uniqueness for classification type problems or lower bound on for regression type problems. Defaults to 0.50.

Methods

<i>name</i>	Returns a name describing the data check.
<i>uniqueness_score</i>	This function calculates a uniqueness score for the provided field. NaN values are
<i>validate</i>	Checks if there are any columns in the input that are too unique in the case of classification

name (*cls*)

Returns a name describing the data check.

static uniqueness_score (*col*)

This function calculates a uniqueness score for the provided field. NaN values are not considered as unique values in the calculation.

Based on the Herfindahl–Hirschman Index.

Parameters *col* (*pd.Series*) – Feature values.

Returns Uniqueness score.

Return type (*float*)

validate (*self, X, y=None*)

Checks if there are any columns in the input that are too unique in the case of classification problems or not unique enough in the case of regression problems.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Ignored. Defaults to None.

Returns

dict with a DataCheckWarning if there are any too unique or not unique enough columns.

Return type *dict*

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'regression_unique_enough': [float(x) for x in range(100)],
...     'regression_not_unique_enough': [float(1) for x in range(100)]
... })
>>> uniqueness_check = UniquenessDataCheck(problem_type="regression",
↳threshold=0.8)
>>> assert uniqueness_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Input columns (regression_not_unique_
↳enough) for regression problem type are not unique enough.",
...                     "data_check_name": "UniquenessDataCheck",
...                     "level": "warning",
...                     "code": "NOT_UNIQUE_ENOUGH",
```

(continues on next page)

(continued from previous page)

```

...         "details": {"column": "regression_not_unique_enough",
↪ 'uniqueness_score': 0.0}}},
...         "actions": [{"code": "DROP_COL",
...         "metadata": {"column": "regression_not_unique_enough"}}}]

```

```

evalml.data_checks.uniqueness_data_check.warning_not_unique_enough = Input columns ({} for {}
evalml.data_checks.uniqueness_data_check.warning_too_unique = Input columns ({} for {} pro

```

utils

Module Contents

Classes Summary

<i>EmptyDataChecks</i>	A collection of data checks.
------------------------	------------------------------

Contents

class evalml.data_checks.utils.**EmptyDataChecks** (*data_checks=None*)

A collection of data checks.

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self, X, y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series, np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

Package Contents

Classes Summary

<i>ClassImbalanceDataCheck</i>	Check if any of the target labels are imbalanced, or if the number of values for each target are below 2 times the number of CV folds. Use for classification problems.
<i>DataCheck</i>	Base class for all data checks. Data checks are a set of heuristics used to determine if there are problems with input data.
<i>DataCheckAction</i>	A recommended action returned by a DataCheck.
<i>DataCheckActionCode</i>	Enum for data check action code.
<i>DataCheckError</i>	DataCheckMessage subclass for errors returned by data checks.
<i>DataCheckMessage</i>	Base class for a message returned by a DataCheck, tagged by name.
<i>DataCheckMessageCode</i>	Enum for data check message code.
<i>DataCheckMessageType</i>	Enum for type of data check message: WARNING or ERROR.
<i>DataChecks</i>	A collection of data checks.
<i>DataCheckWarning</i>	DataCheckMessage subclass for warnings returned by data checks.
<i>DateTimeFormatDataCheck</i>	Checks if the datetime column has equally spaced intervals and is monotonically increasing or decreasing in order
<i>DateTimeNaNDataCheck</i>	Checks each column in the input for datetime features and will issue an error if NaN values are present.
<i>DefaultDataChecks</i>	A collection of basic data checks that is used by AutoML by default.
<i>EmptyDataChecks</i>	A collection of data checks.
<i>HighlyNullDataCheck</i>	Checks if there are any highly-null columns and rows in the input.
<i>IDColumnsDataCheck</i>	Check if any of the features are likely to be ID columns.
<i>InvalidTargetDataCheck</i>	Checks if the target data contains missing or invalid values.
<i>MulticollinearityDataCheck</i>	Check if any set features are likely to be multicollinear.
<i>NaturalLanguageNaNDataCheck</i>	Checks each column in the input for natural language features and will issue an error if NaN values are present.
<i>NoVarianceDataCheck</i>	Check if the target or any of the features have no variance.
<i>OutliersDataCheck</i>	Checks if there are any outliers in input data by using IQR to determine score anomalies. Columns with score anomalies are considered to contain outliers.
<i>SparsityDataCheck</i>	Checks if there are any columns with sparsely populated values in the input.
<i>TargetDistributionDataCheck</i>	Checks if the target data contains certain distributions that may need to be transformed prior training to
<i>TargetLeakageDataCheck</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

continues on next page

Table 135 – continued from previous page

<i>UniquenessDataCheck</i>	Checks if there are any columns in the input that are either too unique for classification problems
----------------------------	---

Contents

class evalml.data_checks.**ClassImbalanceDataCheck** (*threshold=0.1*, *min_samples=100*, *num_cv_folds=3*)

Check if any of the target labels are imbalanced, or if the number of values for each target are below 2 times the number of CV folds. Use for classification problems.

Parameters

- **threshold** (*float*) – The minimum threshold allowed for class imbalance before a warning is raised. This threshold is calculated by comparing the number of samples in each class to the sum of samples in that class and the majority class. For example, a multi-class case with [900, 900, 100] samples per classes 0, 1, and 2, respectively, would have a 0.10 threshold for class 2 ($100 / (900 + 100)$). Defaults to 0.10.
- **min_samples** (*int*) – The minimum number of samples per accepted class. If the minority class is both below the threshold and min_samples, then we consider this severely imbalanced. Must be greater than 0. Defaults to 100.
- **num_cv_folds** (*int*) – The number of cross-validation folds. Must be positive. Choose 0 to ignore this warning. Defaults to 3.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if any target labels are imbalanced beyond a threshold for binary and multiclass problems

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Checks if any target labels are imbalanced beyond a threshold for binary and multiclass problems

Ignores NaN values in target labels if they appear.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features. Ignored.
- **y** (*pd.Series*, *np.ndarray*) – Target labels to check for imbalanced data.

Returns

Dictionary with DataCheckWarnings if imbalance in classes is less than the threshold,
and DataCheckErrors if the number of values for each target is below $2 * \text{num_cv_folds}$.

Return type dict

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame()
>>> y = pd.Series([0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
>>> target_check = ClassImbalanceDataCheck(threshold=0.10)
>>> assert target_check.validate(X, y) == {"errors": [{"message": "The number_
↳ of instances of these targets is less than 2 * the number of cross folds =_
↳ 6 instances: [0]"},
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "error",
...                                     "code": "CLASS_
↳ IMBALANCE_BELOW_FOLDS",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     "warnings": [{"message": "The_
↳ following labels fall below 10% of the target: [0]"},
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "warning",
...                                     "code": "CLASS_
↳ IMBALANCE_BELOW_THRESHOLD",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     {"message": "The_
↳ following labels in the target have severe class imbalance because they_
↳ fall under 10% of the target and have less than 100 samples: [0]"},
...                                     "data_check_name":
↳ "ClassImbalanceDataCheck",
...                                     "level": "warning",
...                                     "code": "CLASS_
↳ IMBALANCE_SEVERE",
...                                     "details": {"target_
↳ values": [0]}}],
...                                     "actions": []}
```

class evalml.data_checks.DataCheck

Base class for all data checks. Data checks are a set of heuristics used to determine if there are problems with input data.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Inspects and validates the input data, runs any necessary calculations or algorithms, and returns a list of warnings and errors if applicable.

name (*cls*)

Returns a name describing the data check.

abstract validate (*self*, *X*, *y=None*)

Inspects and validates the input data, runs any necessary calculations or algorithms, and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*) – The input data of shape [n_samples, n_features]

- **y** (*pd.Series, optional*) – The target data of length [n_samples]

Returns Dictionary of DataCheckError and DataCheckWarning messages

Return type dict (*DataCheckMessage*)

class evalml.data_checks.**DataCheckAction** (*action_code, metadata=None*)

A recommended action returned by a DataCheck.

Parameters

- **action_code** (*DataCheckActionCode*) – Action code associated with the action.
- **metadata** (*dict, optional*) – Additional useful information associated with the action. Defaults to None.

Methods

to_dict

to_dict (*self*)

class evalml.data_checks.**DataCheckActionCode**

Enum for data check action code.

Attributes

DROP_COL	Action code for dropping a column.
DROP_ROWS	Action code for dropping rows.
IM- PUTE_COL	Action code for imputing a column.
TRANS- FORM_TARGET	Action code for transforming the target data.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

class evalml.data_checks.**DataCheckError** (*message, data_check_name, message_code=None, details=None*)

DataCheckMessage subclass for errors returned by data checks.

Attributes

mes- sage_type	DataCheckMessageType.ERROR
---------------------------	----------------------------

Methods

`to_dict`

`to_dict (self)`

class evalml.data_checks.**DataCheckMessage** (*message*, *data_check_name*, *message_code=None*, *details=None*)

Base class for a message returned by a DataCheck, tagged by name.

Parameters

- **message** (*str*) – Message string
- **data_check_name** (*str*) – Name of data check
- **message_code** (`DataCheckMessageCode`) – Message code associated with message. Defaults to None.
- **details** (*dict*) – Additional useful information associated with the message. Defaults to None.

Attributes

message_type	None
---------------------	------

Methods

`to_dict`

`to_dict (self)`

class evalml.data_checks.**DataCheckMessageCode**

Enum for data check message code.

Attributes

CLASS_IMBALANCE_BELOW_FOLDS	Message code for when number of values for each target is below 2 * number of CV folds.
CLASS_IMBALANCE_BELOW_THRESHOLD	Message code for when number of classes is less than the threshold.
CLASS_IMBALANCE_SEVERE	Message code for when balance in classes is less than the threshold and minimum class is less than minimum number of accepted samples.
DATE-TIME_HAS_NAN	Message code for when input datetime columns contain NaN values.
DATE-TIME_HAS_UNEVEN_INTERVALS	Message code for when the datetime values have uneven intervals.
DATE-TIME_INFORMATION_NOT_FOUND	Message code for when datetime information can not be found or is in an unaccepted format.
DATE-TIME_IS_NOT_MONOTONIC	Message code for when the datetime values are not monotonically increasing.
HAS_ID_COLUMNS	Message code for data that has ID columns.
HAS_OUTLIERS	Message code for when outliers are detected.
HIGH_VARIANCE	Message code for when high variance is detected for cross-validation.
HIGHLY_NULL_COLUMNS	Message code for highly null columns.

continues on next page

Table 142 – continued from previous page

HIGHLY_NULL_ROWS	Message code for highly null rows.
IS_MULTICOLLINEAR	Message code for when data is potentially multicollinear.
MIS-MATCHED_INDICES	Message code for when input target and features have mismatched indices.
MIS-MATCHED_INDICES_ORDER	Message code for when input target and features have mismatched indices order. The two have the same index values, but shuffled.
MIS-MATCHED_LENGTHS	Message code for when input target and features have different lengths.
NATURAL_LANGUAGE_HAS_NAN	Message code for when input natural language columns contain NaN values.
NO_VARIANCE	Message code for when data has no variance (1 unique value).
NO_VARIANCE_WITH_NULL	Message code for when data has one unique value and NaN values.
NOT_UNIQUE_ENOUGH	Message code for when data does not possess enough unique values.
TARGET_BINARY_NOT_TWO_UNIQUE_VALUES	Message code for target data for a binary classification problem that does not have two unique values.
TARGET_HAS_NULL	Message code for target data that has null values.
TARGET_INCOMPATIBLE_OBJECTIVE	Message code for target data that has incompatible values for the specified objective
TARGET_IS_EMPTY_OR_FULLY_NULL	Message code for target data that is empty or has all null values.
TARGET_IS_NONE	Message code for when target is None.
TARGET_LEAKAGE	Message code for when target leakage is detected.
TARGET_LOGNORMAL_DISTRIBUTION	Message code for target data with a lognormal distribution.
TARGET_MULTICLASS_HIGH_UNIQUE_CLASS	Message code for target data for a multi classification problem that has an abnormally large number of target values.
TARGET_MULTICLASS_NOT_ENOUGH_CLASSES	Message code for target data for a multi classification problem that does not have more than the minimum number of classes.
TARGET_MULTICLASS_NOT_TWO_EXAMPLES_PER_CLASS	Message code for target data for a multi classification problem that does not have two examples per class.
TARGET_UNSUPPORTED_PROBLEM_TYPE	Message code for target data that is being checked against an unsupported problem type.
TARGET_UNSUPPORTED_TYPE	Message code for target data that is of an unsupported type.
TOO_SPARSE	Message code for when multiclass data has values that are too sparsely populated.
TOO_UNIQUE	Message code for when data possesses too many unique values.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)
The name of the Enum member.

value (*self*)
The value of the Enum member.

```
class evalml.data_checks.DataCheckMessageType
```

Enum for type of data check message: WARNING or ERROR.

Attributes

ERROR	Error message returned by a data check.
WARNING	Warning message returned by a data check.

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

class evalml.data_checks.**DataChecks** (*data_checks=None, data_check_params=None*)

A collection of data checks.

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self, X, y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series, np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

class evalml.data_checks.**DataCheckWarning** (*message, data_check_name, message_code=None, details=None*)

DataCheckMessage subclass for warnings returned by data checks.

Attributes

message_type	DataCheckMessageType.WARNING
---------------------	------------------------------

Methods

<i>to_dict</i>

to_dict (*self*)

class evalml.data_checks.**DateTimeFormatDataCheck** (*datetime_column='index'*)

Checks if the datetime column has equally spaced intervals and is monotonically increasing or decreasing in order to be supported by time series estimators.

Parameters *datetime_column* (*str*, *int*) – The name of the datetime column. If the datetime values are in the index, then pass “index”.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if the target data has equal intervals and is sorted.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Checks if the target data has equal intervals and is sorted.

Parameters

- *X* (*pd.DataFrame*, *np.ndarray*) – Features.
- *y* (*pd.Series*, *np.ndarray*) – Target data.

Returns List with DataCheckErrors if unequal intervals are found in the datetime column.

Return type dict (*DataCheckError*)

Example

```
>>> from pandas as pd
>>> X = pd.DataFrame(pd.date_range("January 1, 2021", periods=8), columns=[
↳ "dates"])
>>> y = pd.Series([1, 2, 4, 2, 1, 2, 3, 1])
>>> X.iloc[7] = "January 9, 2021"
>>> datetime_format_check = DateTimeFormatDataCheck()
>>> assert datetime_format_check.validate(X, y) == {
...     "errors": [{"message": "No frequency could be detected in dates,
↳ possibly due to uneven intervals.",
...                 "data_check_name": "EqualIntervalDataCheck",
...                 "level": "error",
...                 "code": "DATETIME_HAS_UNEVEN_INTERVALS",
...                 "details": {}}],
...     "warnings": [],
...     "actions": []}
```

class evalml.data_checks.**DateTimeNaNDataCheck**

Checks each column in the input for datetime features and will issue an error if NaN values are present.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if any datetime columns contain NaN values.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y=None*)

Checks if any datetime columns contain NaN values.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features.
- **y** (*pd.Series*, *np.ndarray*) – Ignored. Defaults to None.

Returns dict with a `DataCheckError` if NaN values are present in datetime columns.

Return type dict

Example

```
>>> import pandas as pd
>>> import woodwork as ww
>>> import numpy as np
>>> dates = np.arange(np.datetime64('2017-01-01'), np.datetime64('2017-01-08
↳'))
>>> dates[0] = np.datetime64('NaT')
>>> df = pd.DataFrame(dates, columns=['index'])
>>> df.ww.init()
>>> dt_nan_check = DateTimeNaNDataCheck()
>>> assert dt_nan_check.validate(df) == {"warnings": [],
...                                     "actions": [],
...                                     "errors": [
↳[DataCheckError(message='Input datetime column(s) (index) contains NaN
↳values. Please impute NaN values or drop these rows or columns.',
...                                     data_
↳check_name=DateTimeNaNDataCheck.name,
...                                     ]
↳message_code=DataCheckMessageCode.DATETIME_HAS_NAN,
...                                     ]
↳details={"columns": 'index'}).to_dict()]}
```

class evalml.data_checks.**DefaultDataChecks** (*problem_type*, *objective*, *n_splits=3*, *date-time_column=None*)

A collection of basic data checks that is used by AutoML by default. Includes:

- *HighlyNullDataCheck*
- *HighlyNullRowsDataCheck*
- *IDColumnsDataCheck*
- *TargetLeakageDataCheck*
- *InvalidTargetDataCheck*
- *NoVarianceDataCheck*
- *ClassImbalanceDataCheck* (for classification problem types)
- *DateTimeNaNDataCheck*
- *NaturalLanguageNaNDataCheck*
- *TargetDistributionDataCheck* (for regression problem types)
- *DateTimeFormatDataCheck* (for time series problem types)

Parameters

- **problem_type** (*str*) – The problem type that is being validated. Can be regression, binary, or multiclass.
- **objective** (*str* or *ObjectiveBase*) – Name or instance of the objective class.
- **n_splits** (*int*) – The number of splits as determined by the data splitter being used. Defaults to 3.
- **datetime_column** (*str*) – The name of the column containing datetime information to be used for time series problems.
- **to "index" indicating that the datetime information is in the index of X or y. (Default)** –

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self*, *X*, *y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

class evalml.data_checks.**EmptyDataChecks** (*data_checks=None*)

A collection of data checks.

Methods

<i>validate</i>	Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.
-----------------	--

validate (*self*, *X*, *y=None*)

Inspects and validates the input data against data checks and returns a list of warnings and errors if applicable.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*) – The target data of length [n_samples]

Returns Dictionary containing DataCheckMessage objects

Return type dict

class evalml.data_checks.**HighlyNullDataCheck** (*pct_null_col_threshold=0.95*,
pct_null_row_threshold=0.95)

Checks if there are any highly-null columns and rows in the input.

Parameters

- **pct_null_col_threshold** (*float*) – If the percentage of NaN values in an input feature exceeds this amount, that column will be considered highly-null. Defaults to 0.95.
- **pct_null_row_threshold** (*float*) – If the percentage of NaN values in an input row exceeds this amount, that row will be considered highly-null. Defaults to 0.95.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Checks if there are any highly-null columns or rows in the input.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Checks if there are any highly-null columns or rows in the input.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Ignored.

Returns dict with a DataCheckWarning if there are any highly-null columns or rows.

Return type dict

Example

```
>>> import pandas as pd
>>> class SeriesWrap():
...     def __init__(self, series):
...         self.series = series
...
...     def __eq__(self, series_2):
...         return all(self.series.eq(series_2.series))
...
>>> df = pd.DataFrame({
...     'lots_of_null': [None, None, None, None, 5],
...     'no_null': [1, 2, 3, 4, 5]
... })
>>> null_check = HighlyNullDataCheck(pct_null_col_threshold=0.50, pct_null_
↪row_threshold=0.50)
>>> validation_results = null_check.validate(df)
>>> validation_results['warnings'][0]['details']['pct_null_cols'] =
↪SeriesWrap(validation_results['warnings'][0]['details']['pct_null_cols'])
>>> highly_null_rows = SeriesWrap(pd.Series([0.5, 0.5, 0.5, 0.5]))
>>> assert validation_results == {
...     "errors": [],
...     "warnings": [{"message": "4 out of 5 rows are more than 50.0% null",
...         "data_check_name": "HighlyNullDataCheck",
...         "level": "warning",
...         "code": "HIGHLY_NULL_ROWS",
...         "details": {"pct_null_cols": highly_null_rows}},
...     {"message": "Column 'lots_of_null' is 50.0% or more null
↪",
...         "data_check_name": "HighlyNullDataCheck",
```

(continues on next page)

(continued from previous page)

```

...         "level": "warning",
...         "code": "HIGHLY_NULL_COLS",
...         "details": {"column": "lots_of_null", "pct_null_rows":
↪0.8}}},
...     "actions": [{"code": "DROP_ROWS", "metadata": {"rows": [0, 1, 2, 3]}},
...                 {"code": "DROP_COL", "metadata": {"column": "lots_of_null"}
↪}]

```

class evalml.data_checks.IDColumnsDataCheck (*id_threshold=1.0*)

Check if any of the features are likely to be ID columns.

Parameters *id_threshold* (*float*) – The probability threshold to be considered an ID column.
Defaults to 1.0.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if any of the features are likely to be ID columns. Currently performs these simple checks:

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Check if any of the features are likely to be ID columns. Currently performs these simple checks:

- column name is “id”
- column name ends in “_id”
- column contains all unique values (and is categorical / integer type)

Parameters *X* (*pd.DataFrame, np.ndarray*) – The input features to check

Returns A dictionary of features with column name or index and their probability of being ID columns

Return type dict

Example

```

>>> import pandas as pd
>>> df = pd.DataFrame({
...     'df_id': [0, 1, 2, 3, 4],
...     'x': [10, 42, 31, 51, 61],
...     'y': [42, 54, 12, 64, 12]
... })
>>> id_col_check = IDColumnsDataCheck()
>>> assert id_col_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Column 'df_id' is 100.0% or more likely to
↪be an ID column",
...                     "data_check_name": "IDColumnsDataCheck",
...                     "level": "warning",
...                     "code": "HAS_ID_COLUMN",
...                     "details": {"column": "df_id"}}}],

```

(continues on next page)

(continued from previous page)

```
...     "actions": [{"code": "DROP_COL",
...                  "metadata": {"column": "df_id"}}}]}
```

class evalml.data_checks.InvalidTargetDataCheck (*problem_type*, *objective*, *n_unique=100*)

Checks if the target data contains missing or invalid values.

Parameters

- **problem_type** (*str* or *ProblemTypes*) – The specific problem type to data check for. e.g. ‘binary’, ‘multiclass’, ‘regression’, ‘time series regression’
- **objective** (*str* or *ObjectiveBase*) – Name or instance of the objective class.
- **n_unique** (*int*) – Number of unique target values to store when problem type is binary and target incorrectly has more than 2 unique values. Non-negative integer. If None, stores all unique values. Defaults to 100.

Attributes

multi-class_continuous_threshold	0.05
----------------------------------	------

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if the target data contains missing or invalid values.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Checks if the target data contains missing or invalid values.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features. Ignored.
- **y** (*pd.Series*, *np.ndarray*) – Target data to check for invalid values.

Returns List with DataCheckErrors if any invalid values are found in the target data.

Return type dict (*DataCheckError*)

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame({"col": [1, 2, 3, 1]})
>>> y = pd.Series([0, 1, None, None])
>>> target_check = InvalidTargetDataCheck('binary', 'Log Loss Binary')
>>> assert target_check.validate(X, y) == {
...     "errors": [{"message": "2 row(s) (50.0%) of target values are null",
...                  "data_check_name": "InvalidTargetDataCheck",
...                  "level": "error",
...                  "code": "TARGET_HAS_NULL",
```

(continues on next page)

(continued from previous page)

```

...         "details": {"num_null_rows": 2, "pct_null_rows": 50}}],
...         "warnings": [],
...         "actions": [{"code": 'IMPUTE_COL', 'metadata': {'column': None,
↪ 'impute_strategy': 'most_frequent', 'is_target': True}}]}

```

class evalml.data_checks.MulticollinearityDataCheck (*threshold=0.9*)

Check if any set features are likely to be multicollinear.

Parameters *threshold* (*float*) – The threshold to be considered. Defaults to 0.9.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if any set of features are likely to be multicollinear.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Check if any set of features are likely to be multicollinear.

Parameters

- **X** (*pd.DataFrame*) – The input features to check.
- **y** (*pd.Series*) – The target. Ignored.

Returns dict with a DataCheckWarning if there are any potentially multicollinear columns.

Return type dict

Example

```

>>> import pandas as pd
>>> col = pd.Series([1, 0, 2, 3, 4])
>>> X = pd.DataFrame({"col_1": col, "col_2": col * 3})
>>> y = pd.Series([1, 0, 0, 1, 0])
>>> multicollinearity_check = MulticollinearityDataCheck(threshold=0.8)
>>> assert multicollinearity_check.validate(X, y) == {
...     "errors": [],
...     "warnings": [{"message": "Columns are likely to be correlated: [('col_
↪ 1', 'col_2')]",
...     "data_check_name": "MulticollinearityDataCheck",
...     "level": "warning",
...     "code": "IS_MULTICOLLINEAR",
...     "details": {'columns': [('col_1', 'col_2')]}},
...     "actions": []}

```

class evalml.data_checks.NaturalLanguageNaNDataCheck

Checks each column in the input for natural language features and will issue an error if NaN values are present.

Methods

<i>name</i>	Returns a name describing the data check.
	continues on next page

Table 155 – continued from previous page

<code>validate</code>	Checks if any natural language columns contain NaN values.
-----------------------	--

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y=None*)

Checks if any natural language columns contain NaN values.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – Features.
- **y** (*pd.Series, np.ndarray*) – Ignored. Defaults to None.

Returns dict with a DataCheckError if NaN values are present in natural language columns.

Return type dict

Example

```
>>> import pandas as pd
>>> import woodwork as ww
>>> import numpy as np
>>> data = pd.DataFrame()
>>> data['A'] = [None, "string_that_is_long_enough_for_natural_language"]
>>> data['B'] = ['string_that_is_long_enough_for_natural_language', 'string_
↳ that_is_long_enough_for_natural_language']
>>> data['C'] = np.random.randint(0, 3, size=len(data))
>>> data.ww.init(logical_types={'A': 'NaturalLanguage', 'B': 'NaturalLanguage
↳ '})
>>> nl_nan_check = NaturalLanguageNaNDataCheck()
>>> assert nl_nan_check.validate(data) == {
...     "warnings": [],
...     "actions": [],
...     "errors": [DataCheckError(message='Input natural language_
↳ column(s) (A) contains NaN values. Please impute NaN values or drop these_
↳ rows or columns.',
...                               data_check_name=NaturalLanguageNaNDataCheck.name,
...                               message_code=DataCheckMessageCode.NATURAL_LANGUAGE_
↳ HAS_NAN,
...                               details={"columns": 'A'}).to_dict()]
... }
```

class evalml.data_checks.NoVarianceDataCheck (*count_nan_as_value=False*)

Check if the target or any of the features have no variance.

Parameters **count_nan_as_value** (*bool*) – If True, missing values will be counted as their own unique value. Additionally, if true, will return a DataCheckWarning instead of an error if the feature has mostly missing data and only one unique value. Defaults to False.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Check if the target or any of the features have no variance (1 unique value).

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y*)

Check if the target or any of the features have no variance (1 unique value).

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input features.
- **y** (*pd.Series*, *np.ndarray*) – The target data.

Returns dict of warnings/errors corresponding to features or target with no variance.

Return type dict

class evalml.data_checks.OutliersDataCheck

Checks if there are any outliers in input data by using IQR to determine score anomalies. Columns with score anomalies are considered to contain outliers.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Checks if there are any outliers in a dataframe by using IQR to determine column anomalies. Column with anomalies are considered to contain outliers.

name (*cls*)

Returns a name describing the data check.

validate (*self*, *X*, *y=None*)

Checks if there are any outliers in a dataframe by using IQR to determine column anomalies. Column with anomalies are considered to contain outliers.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features
- **y** (*pd.Series*, *np.ndarray*) – Ignored.

Returns A dictionary with warnings if any columns have outliers.

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'x': [1, 2, 3, 4, 5],
...     'y': [6, 7, 8, 9, 10],
...     'z': [-1, -2, -3, -1201, -4]
... })
>>> outliers_check = OutliersDataCheck()
>>> assert outliers_check.validate(df) == {
...     "warnings": [{"message": "Column(s) 'z' are likely to have outlier",
... data."},
...
...                 "data_check_name": "OutliersDataCheck",
...                 "level": "warning",
...                 "code": "HAS_OUTLIERS",
...                 "details": {"columns": ["z"]}
... }
```

(continues on next page)

(continued from previous page)

```
...     "errors": [],
...     "actions": []}
```

class evalml.data_checks.**SparsityDataCheck** (*problem_type*, *threshold*,
unique_count_threshold=10)

Checks if there are any columns with sparsely populated values in the input.

Parameters

- **problem_type** (*str* or *ProblemTypes*) – The specific problem type to data check for. ‘multiclass’ or ‘time series multiclass’ is the only accepted problem type.
- **threshold** (*float*) – The threshold value, or percentage of each column’s unique values, below which, a column exhibits sparsity. Should be between 0 and 1.
- **unique_count_threshold** (*int*) – The minimum number of times a unique value has to be present in a column to not be considered “sparse.” Defaults to 10.

Methods

<i>name</i>	Returns a name describing the data check.
<i>sparsity_score</i>	This function calculates a sparsity score for the given value counts by calculating the percentage of
<i>validate</i>	Calculates what percentage of each column’s unique values exceed the count threshold and compare

name (*cls*)

Returns a name describing the data check.

static sparsity_score (*col*, *count_threshold=10*)

This function calculates a sparsity score for the given value counts by calculating the percentage of unique values that exceed the count_threshold.

Parameters

- **col** (*pd.Series*) – Feature values.
- **count_threshold** (*int*) – The number of instances below which a value is considered sparse. Default is 10.

Returns Sparsity score, or the percentage of the unique values that exceed count_threshold.

Return type (float)

validate (*self*, *X*, *y=None*)

Calculates what percentage of each column’s unique values exceed the count threshold and compare that percentage to the sparsity threshold stored in the class instance.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features.
- **y** (*pd.Series*, *np.ndarray*) – Ignored.

Returns dict with a DataCheckWarning if there are any sparse columns.

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'sparse': [float(x) for x in range(100)],
...     'not_sparse': [float(1) for x in range(100)]
... })
>>> sparsity_check = SparsityDataCheck(problem_type="multiclass", threshold=0.
↳5, unique_count_threshold=10)
>>> assert sparsity_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Input columns (sparse) for multiclass_
↳problem type are too sparse.",
...                     "data_check_name": "SparsityDataCheck",
...                     "level": "warning",
...                     "code": "TOO_SPARSE",
...                     "details": {"column": "sparse", 'sparsity_score': 0.0}}
↳],
...     "actions": [{"code": "DROP_COL",
...                   "metadata": {"column": "sparse"}}}]}
```

class evalml.data_checks.TargetDistributionDataCheck

Checks if the target data contains certain distributions that may need to be transformed prior training to improve model performance.

Methods

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	Checks if the target data has a certain distribution.

`name (cls)`

Returns a name describing the data check.

`validate (self, X, y)`

Checks if the target data has a certain distribution.

Parameters

- **X** (`pd.DataFrame`, `np.ndarray`) – Features. Ignored.
- **y** (`pd.Series`, `np.ndarray`) – Target data to check for underlying distributions.

Returns List with `DataCheckErrors` if certain distributions are found in the target data.

Return type dict (`DataCheckError`)

Example

```
>>> from scipy.stats import lognorm
>>> X = None
>>> y = [0.946, 0.972, 1.154, 0.954, 0.969, 1.222, 1.038, 0.999, 0.973, 0.897]
>>> target_check = TargetDistributionDataCheck()
>>> assert target_check.validate(X, y) == {
...     "errors": [],
...     "warnings": [{"message": "Target may have a lognormal distribution.",
...                             "data_check_name": "TargetDistributionDataCheck",
...                             "level": "warning",
```

(continues on next page)

(continued from previous page)

```

...         "code": "TARGET_LOGNORMAL_DISTRIBUTION",
...         "details": {"shapiro-statistic/pvalue": '0.84/0.045'}}],
...     "actions": [{ 'code': 'TRANSFORM_TARGET', 'metadata': { 'column': None,
↪ 'transformation_strategy': 'lognormal', 'is_target': True}}}]

```

class evalml.data_checks.TargetLeakageDataCheck (*pct_corr_threshold=0.95*,
method='mutual')

Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

If *method='mutual'*, this data check uses mutual information and supports all target and feature types. Otherwise, if *method='pearson'*, it uses Pearson correlation and only supports binary with numeric and boolean dtypes. Pearson correlation returns a value in [-1, 1], while mutual information returns a value in [0, 1].

Parameters

- **pct_corr_threshold** (*float*) – The correlation threshold to be considered leakage. Defaults to 0.95.
- **method** (*string*) – The method to determine correlation. Use ‘mutual’ for mutual information, otherwise ‘pearson’ for Pearson correlation. Defaults to ‘mutual’.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

name (*cls*)

Returns a name describing the data check.

validate (*self, X, y*)

Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.

If *method='mutual'*, supports all target and feature types. Otherwise, if *method='pearson'* only supports binary with numeric and boolean dtypes. Pearson correlation returns a value in [-1, 1], while mutual information returns a value in [0, 1].

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input features to check
- **y** (*pd.Series, np.ndarray*) – The target data

Returns dict with a DataCheckWarning if target leakage is detected.

Return type dict (*DataCheckWarning*)

Example

```
>>> import pandas as pd
>>> X = pd.DataFrame({
...     'leak': [10, 42, 31, 51, 61],
...     'x': [42, 54, 12, 64, 12],
...     'y': [13, 5, 13, 74, 24],
... })
>>> y = pd.Series([10, 42, 31, 51, 40])
>>> target_leakage_check = TargetLeakageDataCheck(pct_corr_threshold=0.95)
>>> assert target_leakage_check.validate(X, y) == {
...     "warnings": [{"message": "Column 'leak' is 95.0% or more correlated_
↳ with the target",
...                     "data_check_name": "TargetLeakageDataCheck",
...                     "level": "warning",
...                     "code": "TARGET_LEAKAGE",
...                     "details": {"column": "leak"}}],
...     "errors": [],
...     "actions": [{"code": "DROP_COL",
...                   "metadata": {"column": "leak"}}]}
```

class evalml.data_checks.UniquenessDataCheck (*problem_type, threshold=0.5*)

Checks if there are any columns in the input that are either too unique for classification problems or not unique enough for regression problems.

Parameters

- **problem_type** (*str or ProblemTypes*) – The specific problem type to data check for. e.g. ‘binary’, ‘multiclass’, ‘regression’, ‘time series regression’
- **threshold** (*float*) – The threshold to set as an upper bound on uniqueness for classification type problems or lower bound on for regression type problems. Defaults to 0.50.

Methods

<i>name</i>	Returns a name describing the data check.
<i>uniqueness_score</i>	This function calculates a uniqueness score for the provided field. NaN values are
<i>validate</i>	Checks if there are any columns in the input that are too unique in the case of classification

name (*cls*)

Returns a name describing the data check.

static uniqueness_score (*col*)

This function calculates a uniqueness score for the provided field. NaN values are not considered as unique values in the calculation.

Based on the Herfindahl–Hirschman Index.

Parameters *col* (*pd.Series*) – Feature values.

Returns Uniqueness score.

Return type (float)

validate (*self, X, y=None*)

Checks if there are any columns in the input that are too unique in the case of classification problems or not unique enough in the case of regression problems.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Features.
- **y** (*pd.Series*, *np.ndarray*) – Ignored. Defaults to None.

Returns

dict with a DataCheckWarning if there are any too unique or not unique enough columns.

Return type dict

Example

```
>>> import pandas as pd
>>> df = pd.DataFrame({
...     'regression_unique_enough': [float(x) for x in range(100)],
...     'regression_not_unique_enough': [float(1) for x in range(100)]
... })
>>> uniqueness_check = UniquenessDataCheck(problem_type="regression",
↳ threshold=0.8)
>>> assert uniqueness_check.validate(df) == {
...     "errors": [],
...     "warnings": [{"message": "Input columns (regression_not_unique_
↳ enough) for regression problem type are not unique enough.",
...                     "data_check_name": "UniquenessDataCheck",
...                     "level": "warning",
...                     "code": "NOT_UNIQUE_ENOUGH",
...                     "details": {"column": "regression_not_unique_enough",
↳ 'uniqueness_score': 0.0}}],
...     "actions": [{"code": "DROP_COL",
...                     "metadata": {"column": "regression_not_unique_enough"}}]}
```

Demos**Submodules****breast_cancer****Module Contents****Functions**

load_breast_cancer

Load breast cancer dataset. Binary classification problem.

Contents

`evalml.demos.breast_cancer.load_breast_cancer()`

Load breast cancer dataset. Binary classification problem.

Returns X and y

Return type (pd.DataFrame, pd.Series)

churn

Module Contents

Functions

<code>load_churn</code>	Load credit card fraud dataset.
-------------------------	---------------------------------

Contents

`evalml.demos.churn.load_churn(n_rows=None, verbose=True)`

Load credit card fraud dataset. The fraud dataset can be used for binary classification problems.

Parameters

- **n_rows** (*int*) – Number of rows from the dataset to return
- **verbose** (*bool*) – Whether to print information about features and labels

Returns X and y

Return type (pd.DataFrame, pd.Series)

diabetes

Module Contents

Functions

<code>load_diabetes</code>	Load diabetes dataset. Regression problem
----------------------------	---

Contents

`evalml.demos.diabetes.load_diabetes()`

Load diabetes dataset. Regression problem

Returns X and y

Return type (pd.DataFrame, pd.Series)

fraud

Module Contents

Functions

<code>load_fraud</code>	Load credit card fraud dataset.
-------------------------	---------------------------------

Contents

`evalml.demos.fraud.load_fraud(n_rows=None, verbose=True)`

Load credit card fraud dataset. The fraud dataset can be used for binary classification problems.

Parameters

- **n_rows** (*int*) – Number of rows from the dataset to return
- **verbose** (*bool*) – Whether to print information about features and labels

Returns X and y

Return type (pd.DataFrame, pd.Series)

weather

Module Contents

Functions

<code>load_weather</code>	Load the Australian daily-min-temperatures weather dataset.
---------------------------	---

Contents

`evalml.demos.weather.load_weather()`

Load the Australian daily-min-termperatures weather dataset.

wine

Module Contents

Functions

<code>load_wine</code>	Load wine dataset. Multiclass problem.
------------------------	--

Contents

`evalml.demos.wine.load_wine()`

Load wine dataset. Multiclass problem.

Returns X and y

Return type (pd.DataFrame, pd.Series)

Package Contents

Functions

<code>load_breast_cancer</code>	Load breast cancer dataset. Binary classification problem.
<code>load_churn</code>	Load credit card fraud dataset.
<code>load_diabetes</code>	Load diabetes dataset. Regression problem
<code>load_fraud</code>	Load credit card fraud dataset.
<code>load_weather</code>	Load the Australian daily-min-termperatures weather dataset.
<code>load_wine</code>	Load wine dataset. Multiclass problem.

Contents

`evalml.demos.load_breast_cancer()`

Load breast cancer dataset. Binary classification problem.

Returns X and y

Return type (pd.DataFrame, pd.Series)

`evalml.demos.load_churn(n_rows=None, verbose=True)`

Load credit card fraud dataset. The fraud dataset can be used for binary classification problems.

Parameters

- **n_rows** (*int*) – Number of rows from the dataset to return
- **verbose** (*bool*) – Whether to print information about features and labels

Returns X and y

Return type (pd.DataFrame, pd.Series)

```
evalml.demos.load_diabetes()
```

Load diabetes dataset. Regression problem

Returns X and y

Return type (pd.DataFrame, pd.Series)

```
evalml.demos.load_fraud(n_rows=None, verbose=True)
```

Load credit card fraud dataset. The fraud dataset can be used for binary classification problems.

Parameters

- **n_rows** (*int*) – Number of rows from the dataset to return
- **verbose** (*bool*) – Whether to print information about features and labels

Returns X and y

Return type (pd.DataFrame, pd.Series)

```
evalml.demos.load_weather()
```

Load the Australian daily-min-termperatures weather dataset.

```
evalml.demos.load_wine()
```

Load wine dataset. Multiclass problem.

Returns X and y

Return type (pd.DataFrame, pd.Series)

Exceptions

Submodules

exceptions

Module Contents

Classes Summary

PartialDependenceErrorCode

Enum identifying the type of error encountered in partial dependence.

Exceptions Summary

Contents

exception `evalml.exceptions.exceptions.AutoMLSearchException`

Exception raised when all pipelines in an automl batch return a score of NaN for the primary objective.

exception `evalml.exceptions.exceptions.ComponentNotYetFittedError`

An exception to be raised when `predict/predict_proba/transform` is called on a component without fitting first.

exception `evalml.exceptions.exceptions.DataCheckInitError`

Exception raised when a data check can't initialize with the parameters given.

exception `evalml.exceptions.exceptions.EnsembleMissingPipelinesError`

An exception raised when an ensemble is missing *estimators* (list) as a parameter.

exception `evalml.exceptions.exceptions.MethodPropertyNotFoundError`

Exception to raise when a class does not have an expected method or property.

exception `evalml.exceptions.exceptions.MissingComponentError`

An exception raised when a component is not found in `all_components()`

exception `evalml.exceptions.exceptions.NoPositiveLabelException`

Exception when a particular classification label for the 'positive' class cannot be found in the column index or unique values

exception `evalml.exceptions.exceptions.NullsInColumnWarning`

Warning thrown when there are null values in the column of interest

exception `evalml.exceptions.exceptions.ObjectiveCreationError`

Exception when `get_objective` tries to instantiate an objective and required args are not provided.

exception `evalml.exceptions.exceptions.ObjectiveNotFoundError`

Exception to raise when specified objective does not exist.

exception `evalml.exceptions.exceptions.ParameterNotUsedWarning` (*components*)

Warning thrown when a pipeline parameter isn't used in a defined pipeline's component graph during initialization.

exception `evalml.exceptions.exceptions.PartialDependenceError` (*message, code*)

Exception raised for all errors that partial dependence can raise.

class `evalml.exceptions.exceptions.PartialDependenceErrorCode`

Enum identifying the type of error encountered in partial dependence.

Attributes

ALL_OTHER_ERRORS	errors
COMPUTED_PERCENTILES_TOO_CLOSE	computed_percentiles_too_close
FEATURE_IS_ALL_NANS	feature_is_all_nans
FEATURE_IS_MOSTLY_ONE_VALUE	feature_is_mostly_one_value
FEATURES_ARGUMENT_INCORRECT_TYPES	features_argument_incorrect_types
ICE_PLOT_REQUESTED_FOR_TWO_WAY_PLOT	ice_dependence_plot
INVALID_CLASS_LABEL	invalid_class_label_requested_for_plot
INVALID_FEATURE_TYPE	invalid_feature_type
PIPELINE_IS_BASELINE	baseline
TOO_MANY_FEATURES	features
TWO_WAY_REQUESTED_FOR_DATES	dates
UNFITTED_PIPELINE	unfitted_pipeline

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

exception evalml.exceptions.exceptions.**PipelineNotFoundError**

An exception raised when a particular pipeline is not found in automl search results

exception evalml.exceptions.exceptions.**PipelineNotYetFittedError**

An exception to be raised when predict/predict_proba/transform is called on a pipeline without fitting first.

exception evalml.exceptions.exceptions.**PipelineScoreError** (*exceptions*, *scored_successfully*)

An exception raised when a pipeline errors while scoring any objective in a list of objectives.

Parameters

- **exceptions** (*dict*) – A dictionary mapping an objective name (str) to a tuple of the form (exception, traceback). All of the objectives that errored will be stored here.
- **scored_successfully** (*dict*) – A dictionary mapping an objective name (str) to a score value. All of the objectives that did not error will be stored here.

Package Contents

Classes Summary

<i>PartialDependenceErrorCode</i>	Enum identifying the type of error encountered in partial dependence.
-----------------------------------	---

Exceptions Summary

Contents

- exception** `evalml.exceptions.AutoMLSearchException`
Exception raised when all pipelines in an automl batch return a score of NaN for the primary objective.
- exception** `evalml.exceptions.ComponentNotYetFittedError`
An exception to be raised when `predict/predict_proba/transform` is called on a component without fitting first.
- exception** `evalml.exceptions.DataCheckInitError`
Exception raised when a data check can't initialize with the parameters given.
- exception** `evalml.exceptions.EnsembleMissingPipelinesError`
An exception raised when an ensemble is missing *estimators* (list) as a parameter.
- exception** `evalml.exceptions.MethodPropertyNotFoundError`
Exception to raise when a class does not have an expected method or property.
- exception** `evalml.exceptions.MissingComponentError`
An exception raised when a component is not found in `all_components()`
- exception** `evalml.exceptions.NoPositiveLabelException`
Exception when a particular classification label for the 'positive' class cannot be found in the column index or unique values
- exception** `evalml.exceptions.NullsInColumnWarning`
Warning thrown when there are null values in the column of interest
- exception** `evalml.exceptions.ObjectiveCreationError`
Exception when `get_objective` tries to instantiate an objective and required args are not provided.
- exception** `evalml.exceptions.ObjectiveNotFoundError`
Exception to raise when specified objective does not exist.
- exception** `evalml.exceptions.ParameterNotUsedWarning` (*components*)
Warning thrown when a pipeline parameter isn't used in a defined pipeline's component graph during initialization.
- exception** `evalml.exceptions.PartialDependenceError` (*message, code*)
Exception raised for all errors that partial dependence can raise.
- class** `evalml.exceptions.PartialDependenceErrorCode`
Enum identifying the type of error encountered in partial dependence.

Attributes

ALL_OTHER_ERRORS	errors
COMPUTED_PERCENTILES_TOO_CLOSE	computed_percentiles_too_close
FEATURE_IS_ALL_NANS	feature_is_all_nans
FEATURE_IS_MOSTLY_ONE_VALUE	feature_is_mostly_one_value
FEATURES_ARGUMENT_INCORRECT_TYPES	features_argument_incorrect_types
ICE_PLOT_REQUESTED_FOR_TWO_WAY_PLOT	ice_dependence_plot
INVALID_CLASS_LABEL	invalid_class_label_requested_for_plot
INVALID_FEATURE_TYPE	invalid_feature_type
PIPELINE_IS_BASELINE	baseline
TOO_MANY_FEATURES	features
TWO_WAY_REQUESTED_FOR_DATES	dates
UNFITTED_PIPELINE	unfitted_pipeline

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

exception evalml.exceptions.PipelineNotFoundError

An exception raised when a particular pipeline is not found in automl search results

exception evalml.exceptions.PipelineNotYetFittedError

An exception to be raised when predict/predict_proba/transform is called on a pipeline without fitting first.

exception evalml.exceptions.PipelineScoreError (*exceptions, scored_successfully*)

An exception raised when a pipeline errors while scoring any objective in a list of objectives.

Parameters

- **exceptions** (*dict*) – A dictionary mapping an objective name (str) to a tuple of the form (exception, traceback). All of the objectives that errored will be stored here.
- **scored_successfully** (*dict*) – A dictionary mapping an objective name (str) to a score value. All of the objectives that did not error will be stored here.

Model Family

Submodules

model_family

Module Contents

Classes Summary

<i>ModelFamily</i>	Enum for family of machine learning models.
--------------------	---

Contents

class evalml.model_family.model_family.**ModelFamily**

Enum for family of machine learning models.

Attributes

ARIMA	ARIMA model family.
BASELINE	Baseline model family.
CAT-BOOST	CatBoost model family.
DECISION_TREE	Decision Tree model family.
ENSEMBLE	Ensemble model family.
EXTRA_TREES	Extra Trees model family.
K_NEIGHBORS	K Nearest Neighbors model family.
LIGHTGBM	LightGBM model family.
LINEAR_MODEL	Linear model family.
NONE	None
PROPHET	Prophet model family.
RANDOM_FOREST	Random Forest model family.
SVM	SVM model family.
XGBOOST	XGBoost model family.

Methods

<i>is_tree_estimator</i>	Checks whether the estimator's model family uses trees.
<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

is_tree_estimator(*self*)

Checks whether the estimator’s model family uses trees.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

utils

Module Contents

Functions

<i>handle_model_family</i>	Handles <i>model_family</i> by either returning the <i>ModelFamily</i> or converting from a string
----------------------------	--

Contents

`evalml.model_family.utils.handle_model_family(model_family)`

Handles *model_family* by either returning the *ModelFamily* or converting from a string

Parameters *model_family* (*str* or *ModelFamily*) – Model type that needs to be handled

Returns *ModelFamily*

Package Contents

Classes Summary

<i>ModelFamily</i>	Enum for family of machine learning models.
--------------------	---

Functions

<i>handle_model_family</i>	Handles <i>model_family</i> by either returning the <i>ModelFamily</i> or converting from a string
----------------------------	--

Contents

`evalml.model_family.handle_model_family(model_family)`

Handles *model_family* by either returning the *ModelFamily* or converting from a string

Parameters *model_family* (*str* or *ModelFamily*) – Model type that needs to be handled

Returns *ModelFamily*

class `evalml.model_family.ModelFamily`

Enum for family of machine learning models.

Attributes

ARIMA	ARIMA model family.
BASELINE	Baseline model family.
CAT-BOOST	CatBoost model family.
DECISION_TREE	Decision Tree model family.
ENSEMBLE	Ensemble model family.
EXTRA_TREES	Extra Trees model family.
K_NEIGHBORS	K Nearest Neighbors model family.
LIGHTGBM	LightGBM model family.
LINEAR_MODEL	Linear model family.
NONE	None
PROPHET	Prophet model family.
RANDOM_FOREST	Random Forest model family.
SVM	SVM model family.
XGBOOST	XGBoost model family.

Methods

<code>is_tree_estimator</code>	Checks whether the estimator's model family uses trees.
<code>name</code>	The name of the Enum member.
<code>value</code>	The value of the Enum member.

is_tree_estimator (*self*)
Checks whether the estimator's model family uses trees.

name (*self*)
The name of the Enum member.

value (*self*)
The value of the Enum member.

Model Understanding

Subpackages

prediction_explanations

Submodules

explainers

Module Contents

Classes Summary

<code>ExplainPredictionsStage</code>	Generic enumeration.
--------------------------------------	----------------------

Functions

<code>abs_error</code>	Computes the absolute error per data point for regression problems.
<code>cross_entropy</code>	Computes Cross Entropy Loss per data point for classification problems.
<code>explain_predictions</code>	Creates a report summarizing the top contributing features for each data point in the input features.
<code>explain_predictions_best_worst</code>	Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

Attributes Summary

<code>DEFAULT_METRICS</code>

Contents

`evalml.model_understanding.prediction_explanations.explainers.abs_error` (*y_true*, *y_pred*)

Computes the absolute error per data point for regression problems.

Parameters

- **y_true** (*pd.Series*) – True labels.
- **y_pred** (*pd.Series*) – Predicted values.

Returns *np.ndarray*

`evalml.model_understanding.prediction_explanations.explainers.cross_entropy` (*y_true*, *y_pred_proba*)

Computes Cross Entropy Loss per data point for classification problems.

Parameters

- **y_true** (*pd.Series*) – True labels encoded as ints.
- **y_pred_proba** (*pd.DataFrame*) – Predicted probabilities. One column per class.

Returns *np.ndarray*

`evalml.model_understanding.prediction_explanations.explainers.DEFAULT_METRICS`

```

evalml.model_understanding.prediction_explanations.explainers.explain_predictions(pipeline,
in-
put_features,
y,
in-
dices_to_explain,
top_k_features,
in-
clude_shap_values,
in-
clude_expected_value,
output_format="text",
training_data=None,
training_target=None)

```

Creates a report summarizing the top contributing features for each data point in the input features.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Dataframe of input data to evaluate the pipeline on.
- **y** (*pd.Series*) – Labels for the input data.
- **indices_to_explain** (*list(int)*) – List of integer indices to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point. Default is 3.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **include_expected_value** (*bool*) – Whether the expected value should be included in the table. Default is False.
- **output_format** (*str*) – Either “text”, “dict”, or “dataframe”. Default is “text”.
- **training_data** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series, np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features to each prediction for each row of `input_features`. The report will include the feature names, prediction contribution, and SHAP Value (optional).

Raises

- **ValueError** – if `input_features` is empty.
- **ValueError** – if an `output_format` outside of “text”, “dict” or “dataframe” is provided.
- **ValueError** – if the requested index falls outside the `input_features`’s boundaries.

```
evalml.model_understanding.prediction_explanations.explainers.explain_predictions_best_worst
```

Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Input data to evaluate the pipeline on.
- **y_true** (*pd.Series*) – True labels for the input data.
- **num_to_explain** (*int*) – How many of the best, worst, random data points to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **metric** (*callable*) – The metric used to identify the best and worst points in the dataset. Function must accept the true labels and predicted value or probabilities as the only arguments and lower values must be better. By default, this will be the absolute error for regression problems and cross entropy loss for classification problems.
- **output_format** (*str*) – Either “text” or “dict”. Default is “text”.
- **callback** (*callable*) – Function to be called with incremental updates. Has the following parameters: - *progress_stage*: stage of computation - *time_elapsed*: total time in seconds that has elapsed since start of call
- **training_data** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series, np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features for the best/worst predictions in the dataset. For each of the best/worst rows of input_features, the predicted values, true labels, metric value, feature names, prediction contribution, and SHAP Value (optional) will be listed.

Raises

- **ValueError** – if `input_features` does not have more than twice the requested features to explain.
- **ValueError** – if `y_true` and `input_features` have mismatched lengths.
- **ValueError** – if an `output_format` outside of “text”, “dict” or “dataframe” is provided.

class evalml.model_understanding.prediction_explanations.explainers.**ExplainPredictionsStage**

Generic enumeration.

Derive from this class to define new enumerations.

Attributes

COM- PUTE_FEATURE_STAGE	<code>compute_feature_stage</code>
COM- PUTE_SHAP_VALUES_STAGE	<code>compute_shap_value_stage</code>
DONE	<code>done</code>
PRE- DICT_STAGE	<code>predict_stage</code>
PREPRO- CESS- ING_STAGE	<code>preprocessing_stage</code>

Methods

<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

Package Contents**Functions**

<i>explain_predictions</i>	Creates a report summarizing the top contributing features for each data point in the input features.
<i>explain_predictions_best_worst</i>	Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

Contents

```
evalml.model_understanding.prediction_explanations.explain_predictions(pipeline,  
                                in-  
                                put_features,  
                                y,  
                                in-  
                                dices_to_explain,  
                                top_k_features=3,  
                                in-  
                                clude_shap_values=False,  
                                in-  
                                clude_expected_value=False,  
                                out-  
                                put_format='text',  
                                train-  
                                ing_data=None,  
                                train-  
                                ing_target=None)
```

Creates a report summarizing the top contributing features for each data point in the input features.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Dataframe of input data to evaluate the pipeline on.
- **y** (*pd.Series*) – Labels for the input data.
- **indices_to_explain** (*list(int)*) – List of integer indices to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point. Default is 3.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **include_expected_value** (*bool*) – Whether the expected value should be included in the table. Default is False.
- **output_format** (*str*) – Either “text”, “dict”, or “dataframe”. Default is “text”.
- **training_data** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series, np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features to each prediction for each row of `input_features`. The report will include the feature names, prediction contribution, and SHAP Value (optional).

Raises

- **ValueError** – if `input_features` is empty.

- **ValueError** – if an `output_format` outside of “text”, “dict” or “dataframe” is provided.
- **ValueError** – if the requested index falls outside the `input_feature`’s boundaries.

`evalml.model_understanding.prediction_explanations.explain_predictions_best_worst` (*pipeline*, *input_features*, *y_true*, *num_to_explain*, *top_k_features*, *include_shap_values*, *metric=None*, *output_format=None*, *callback=None*, *training_data=None*, *training_target=None*)

Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Input data to evaluate the pipeline on.
- **y_true** (*pd.Series*) – True labels for the input data.
- **num_to_explain** (*int*) – How many of the best, worst, random data points to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **metric** (*callable*) – The metric used to identify the best and worst points in the dataset. Function must accept the true labels and predicted value or probabilities as the only arguments and lower values must be better. By default, this will be the absolute error for regression problems and cross entropy loss for classification problems.
- **output_format** (*str*) – Either “text” or “dict”. Default is “text”.
- **callback** (*callable*) – Function to be called with incremental updates. Has the following parameters: - `progress_stage`: stage of computation - `time_elapsed`: total time in seconds that has elapsed since start of call
- **training_data** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features for the best/worst predictions in the dataset. For each of the best/worst rows of `input_features`, the predicted values, true labels, metric value, feature names, prediction contribution, and SHAP Value (optional) will be listed.

Raises

- **ValueError** – if `input_features` does not have more than twice the requested features to explain.
- **ValueError** – if `y_true` and `input_features` have mismatched lengths.
- **ValueError** – if an `output_format` outside of “text”, “dict” or “dataframe” is provided.

Submodules

force_plots

Module Contents

Functions

<code>force_plot</code>	Function to generate the data required to build a force plot.
<code>graph_force_plot</code>	Function to generate force plots for the desired rows of the training data.

Contents

`evalml.model_understanding.force_plots.force_plot` (*pipeline*, *rows_to_explain*, *training_data*, *y*)

Function to generate the data required to build a force plot.

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to generate the force plot for.
- **rows_to_explain** (*list(int)*) – A list of the indices of the `training_data` to explain.
- **training_data** (*pandas.DataFrame*) – The data used to train the pipeline.
- **y** (*pandas.Series*) – The target data.

Returns

list of dictionaries where each dict contains force plot data. Each dictionary entry represents the explanations for a single row.

For single row binary force plots:

```
[{'malignant': {'expected_value': 0.37, 'feature_names': ['worst concave points', 'worst perimeter', 'worst radius'], 'shap_values': [0.09, 0.09, 0.08], 'plot': AdditiveForceVisualizer}]
```

For two row binary force plots:

```
[{'malignant': {'expected_value': 0.37, 'feature_names': ['worst concave points',
```

```
    'worst perimeter', 'worst radius'], 'shap_values': [0.09, 0.09, 0.08], 'plot': Additive-
    ForceVisualizer},
```

```
    {'malignant': {'expected_value': 0.29, 'feature_names': ['worst concave points',
    'worst perimeter', 'worst radius'], 'shap_values': [0.05, 0.03, 0.02], 'plot': Additive-
    ForceVisualizer}]}
```

Return type list(dict())

Raises

- **TypeError** – if rows_to_explain is not a list.
- **TypeError** – if all values in rows_to_explain aren't integers.

```
evalml.model_understanding.force_plots.graph_force_plot(pipeline, rows_to_explain,
                                                         training_data, y, mat-
                                                         plotlib=False)
```

Function to generate force plots for the desired rows of the training data.

Parameters

- **pipeline** (*PipelineBase*) – The pipeline to generate the force plot for.
- **rows_to_explain** (*list(int)*) – A list of the indices indicating which of the rows of the training_data to explain.
- **training_data** (*pandas.DataFrame*) – The data used to train the pipeline.
- **y** (*pandas.Series*) – The target data for the pipeline.
- **matplotlib** (*bool*) – flag to display the force plot using matplotlib (outside of jupyter) Defaults to False.

Returns

The same as `force_plot()`, but with an additional `key` in each dictionary for the plot.

Return type list(dict(shap.AdditiveForceVisualizer))

graphs

Module Contents

Functions

<code>binary_objective_vs_threshold</code>	Computes objective score as a function of potential binary classification
<code>confusion_matrix</code>	Confusion matrix for binary and multiclass classification.
<code>decision_tree_data_from_estimator</code>	Return data for a fitted tree in a restructured format
<code>decision_tree_data_from_pipeline</code>	Return data for a fitted pipeline with in a restructured format
<code>get_linear_coefficients</code>	Returns a dataframe showing the features with the greatest predictive power for a linear model.
<code>get_prediction_vs_actual_data</code>	Combines <code>y_true</code> and <code>y_pred</code> into a single dataframe and adds a column for outliers. Used in <code>graph_prediction_vs_actual()</code> .

continues on next page

Table 187 – continued from previous page

<code>get_prediction_vs_actual_over_time_data</code>	Get the data needed for the prediction_vs_actual_over_time plot.
<code>graph_binary_objective_vs_threshold</code>	Generates a plot graphing objective score vs. decision thresholds for a fitted binary classification pipeline.
<code>graph_confusion_matrix</code>	Generate and display a confusion matrix plot.
<code>graph_partial_dependence</code>	Create an one-way or two-way partial dependence plot. Passing a single integer or
<code>graph_permutation_importance</code>	Generate a bar graph of the pipeline’s permutation importance.
<code>graph_precision_recall_curve</code>	Generate and display a precision-recall plot.
<code>graph_prediction_vs_actual</code>	Generate a scatter plot comparing the true and predicted values. Used for regression plotting
<code>graph_prediction_vs_actual_over_time</code>	Plot the target values and predictions against time on the x-axis.
<code>graph_roc_curve</code>	Generate and display a Receiver Operating Characteristic (ROC) plot for binary and multiclass classification problems.
<code>graph_t_sne</code>	Plot high dimensional data into lower dimensional space using t-SNE .
<code>normalize_confusion_matrix</code>	Normalizes a confusion matrix.
<code>partial_dependence</code>	Calculates one or two-way partial dependence. If a single integer or
<code>precision_recall_curve</code>	Given labels and binary classifier predicted probabilities, compute and return the data representing a precision-recall curve.
<code>roc_curve</code>	Given labels and classifier predicted probabilities, compute and return the data representing a Receiver Operating Characteristic (ROC) curve. Works with binary or multiclass problems.
<code>t_sne</code>	Get the transformed output after fitting X to the embedded space using t-SNE.
<code>visualize_decision_tree</code>	Generate an image visualizing the decision tree

Contents

`evalml.model_understanding.graphs.binary_objective_vs_threshold`(*pipeline*, *X*, *y*, *objective*, *steps=100*)

Computes objective score as a function of potential binary classification decision thresholds for a fitted binary classification pipeline.

Parameters

- **pipeline** (*BinaryClassificationPipeline obj*) – Fitted binary classification pipeline
- **X** (*pd.DataFrame*) – The input data used to compute objective score
- **y** (*pd.Series*) – The target labels
- **objective** (*ObjectiveBase obj, str*) – Objective used to score
- **steps** (*int*) – Number of intervals to divide and calculate objective score at

Returns DataFrame with thresholds and the corresponding objective score calculated at each threshold

Return type pd.DataFrame

`evalml.model_understanding.graphs.confusion_matrix(y_true, y_predicted, normalize_method='true')`

Confusion matrix for binary and multiclass classification.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier.
- **normalize_method** (*{'true', 'pred', 'all', None}*) – Normalization method to use, if not None. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.

Returns Confusion matrix. The column header represents the predicted labels while row header represents the actual labels.

Return type pd.DataFrame

`evalml.model_understanding.graphs.decision_tree_data_from_estimator(estimator)`

Return data for a fitted tree in a restructured format

Parameters **estimator** (*ComponentBase*) – A fitted DecisionTree-based estimator.

Returns An OrderedDict of OrderedDicts describing a tree structure

Return type OrderedDict

`evalml.model_understanding.graphs.decision_tree_data_from_pipeline(pipeline_)`

Return data for a fitted pipeline with in a restructured format

Parameters **pipeline** (*PipelineBase*) – A pipeline with a DecisionTree-based estimator.

Returns An OrderedDict of OrderedDicts describing a tree structure

Return type OrderedDict

`evalml.model_understanding.graphs.get_linear_coefficients(estimator, features=None)`

Returns a dataframe showing the features with the greatest predictive power for a linear model.

Parameters

- **estimator** (*Estimator*) – Fitted linear model family estimator.
- **features** (*list[str]*) – List of feature names associated with the underlying data.

Returns Displaying the features by importance.

Return type pd.DataFrame

`evalml.model_understanding.graphs.get_prediction_vs_actual_data(y_true, y_pred, outlier_threshold=None)`

Combines y_true and y_pred into a single dataframe and adds a column for outliers. Used in `graph_prediction_vs_actual()`.

Parameters

- **y_true** (*pd.Series*, or *np.ndarray*) – The real target values of the data

- **y_pred** (*pd.Series, or np.ndarray*) – The predicted values outputted by the regression model.
- **outlier_threshold** (*int, float*) – A positive threshold for what is considered an outlier value. This value is compared to the absolute difference between each value of y_true and y_pred. Values within this threshold will be blue, otherwise they will be yellow. Defaults to None

Returns

- *prediction*: Predicted values from regression model.
- *actual*: Real target values.
- *outlier*: Colors indicating which values are in the threshold for what is considered an outlier value.

Return type `pd.DataFrame` with the following columns

```
evalml.model_understanding.graphs.get_prediction_vs_actual_over_time_data(pipeline,
                                                                            X,
                                                                            y,
                                                                            X_train,
                                                                            y_train,
                                                                            dates)
```

Get the data needed for the prediction_vs_actual_over_time plot.

Parameters

- **pipeline** (*TimeSeriesRegressionPipeline*) – Fitted time series regression pipeline.
- **X** (*pd.DataFrame*) – Features used to generate new predictions.
- **y** (*pd.Series*) – Target values to compare predictions against.
- **X_train** (*pd.DataFrame*) – Data the pipeline was trained on.
- **y_train** (*pd.Series*) – Target values for training data.
- **dates** (*pd.Series*) – Dates corresponding to target values and predictions.

Returns `pd.DataFrame`

```
evalml.model_understanding.graphs.graph_binary_objective_vs_threshold(pipeline,
                                                                       X, y,
                                                                       ob-
                                                                       jec-
                                                                       tive,
                                                                       steps=100)
```

Generates a plot graphing objective score vs. decision thresholds for a fitted binary classification pipeline.

Parameters

- **pipeline** (*PipelineBase or subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*) – The input data used to score and compute scores
- **y** (*pd.Series*) – The target labels
- **objective** (*ObjectiveBase obj, str*) – Objective used to score, shown on the y-axis of the graph
- **steps** (*int*) – Number of intervals to divide and calculate objective score at

Returns `plotly.Figure` representing the objective score vs. threshold graph generated

```
evalml.model_understanding.graphs.graph_confusion_matrix(y_true, y_pred, normalize_method='true', title_addition=None)
```

Generate and display a confusion matrix plot.

If `normalize_method` is set, hover text will show raw count, otherwise hover text will show count normalized with method 'true'.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier.
- **normalize_method** (*{'true', 'pred', 'all', None}*) – Normalization method to use, if not None. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.
- **title_addition** (*str* or *None*) – if not None, append to plot title. Defaults to None.

Returns *plotly.Figure* representing the confusion matrix plot generated

```
evalml.model_understanding.graphs.graph_partial_dependence(pipeline, X, features, class_label=None, grid_resolution=100, kind='average')
```

Create an one-way or two-way partial dependence plot. Passing a single integer or string as features will create a one-way partial dependence plot with the feature values plotted against the partial dependence. Passing features a tuple of int/strings will create a two-way partial dependence plot with a contour of feature[0] in the y-axis, feature[1] in the x-axis and the partial dependence in the z-axis.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*, *np.ndarray*) – The input data used to generate a grid of values for feature where partial dependence will be calculated at
- **features** (*int*, *string*, *tuple[int or string]*) – The target feature for which to create the partial dependence plot for. If features is an int, it must be the index of the feature to use. If features is a string, it must be a valid column name in X. If features is a tuple of strings, it must contain valid column int/names in X.
- **class_label** (*string*, *optional*) – Name of class to plot for multiclass problems. If None, will plot the partial dependence for each class. This argument does not change behavior for regression or binary classification pipelines. For binary classification, the partial dependence for the positive label will always be displayed. Defaults to None.
- **grid_resolution** (*int*) – Number of samples of feature(s) for partial dependence plot
- **{'average'}** (*kind*) – Type of partial dependence to plot. 'average' creates a regular partial dependence (PD) graph, 'individual' creates an individual conditional expectation (ICE) plot, and 'both' creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.
- **'individual'** – Type of partial dependence to plot. 'average' creates a regular partial dependence (PD) graph, 'individual' creates an individual conditional expectation (ICE) plot, and 'both' creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.
- **'both' }** – Type of partial dependence to plot. 'average' creates a regular partial dependence (PD) graph, 'individual' creates an individual conditional expectation (ICE) plot, and

‘both’ creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.

Returns figure object containing the partial dependence data for plotting

Return type `plotly.graph_objects.Figure`

Raises

- **PartialDependenceError** – if a graph is requested for a class name that isn’t present in the pipeline.
- **PartialDependenceError** – if an ICE plot is requested for a two-way partial dependence.

```
evalml.model_understanding.graphs.graph_permutation_importance(pipeline, X,  
                                                                y, objective,  
                                                                importance_threshold=0)
```

Generate a bar graph of the pipeline’s permutation importance.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance
- **y** (*pd.Series*) – The target data
- **objective** (*str*, *ObjectiveBase*) – Objective to score on
- **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their respective permutation importance.

```
evalml.model_understanding.graphs.graph_precision_recall_curve(y_true,  
                                                              y_pred_proba,  
                                                              title_addition=None)
```

Generate and display a precision-recall plot.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier, before thresholding has been applied. Note this should be the predicted probability for the “true” label.
- **title_addition** (*str* or *None*) – If not *None*, append to plot title. Default *None*.

Returns `plotly.Figure` representing the precision-recall plot generated

```
evalml.model_understanding.graphs.graph_prediction_vs_actual(y_true, y_pred, outlier_threshold=None)
```

Generate a scatter plot comparing the true and predicted values. Used for regression plotting

Parameters

- **y_true** (*pd.Series*) – The real target values of the data
- **y_pred** (*pd.Series*) – The predicted values outputted by the regression model.

- **outlier_threshold** (*int, float*) – A positive threshold for what is considered an outlier value. This value is compared to the absolute difference between each value of `y_true` and `y_pred`. Values within this threshold will be blue, otherwise they will be yellow. Defaults to None

Returns `plotly.Figure` representing the predicted vs. actual values graph

```
evalml.model_understanding.graphs.graph_prediction_vs_actual_over_time(pipeline,  
                                                                    X, y,  
                                                                    X_train,  
                                                                    y_train,  
                                                                    dates)
```

Plot the target values and predictions against time on the x-axis.

Parameters

- **pipeline** (*TimeSeriesRegressionPipeline*) – Fitted time series regression pipeline.
- **X** (*pd.DataFrame*) – Features used to generate new predictions.
- **y** (*pd.Series*) – Target values to compare predictions against.
- **X_train** (*pd.DataFrame*) – Data the pipeline was trained on.
- **y_train** (*pd.Series*) – Target values for training data.
- **dates** (*pd.Series*) – Dates corresponding to target values and predictions.

Returns Showing the prediction vs actual over time.

Return type `plotly.Figure`

```
evalml.model_understanding.graphs.graph_roc_curve(y_true,      y_pred_proba,      cus-  
                                                    tom_class_names=None,      ti-  
                                                    tle_addition=None)
```

Generate and display a Receiver Operating Characteristic (ROC) plot for binary and multiclass classification problems.

Parameters

- **y_true** (*pd.Series or np.ndarray*) – True labels.
- **y_pred_proba** (*pd.Series or np.ndarray*) – Predictions from a classifier, before thresholding has been applied. Note this should be a one dimensional array with the predicted probability for the “true” label in the binary case.
- **custom_class_labels** (*list or None*) – If not None, custom labels for classes. Default None.
- **title_addition** (*str or None*) – if not None, append to plot title. Default None.

Returns `plotly.Figure` representing the ROC plot generated

```
evalml.model_understanding.graphs.graph_t_sne(X, n_components=2, perplexity=30.0,  
                                              learning_rate=200.0, metric='euclidean',  
                                              marker_line_width=2, marker_size=7,  
                                              **kwargs)
```

Plot high dimensional data into lower dimensional space using t-SNE.

Parameters

- **X** (*np.ndarray, pd.DataFrame*) – Data to be transformed. Must be numeric.
- **n_components** (*int, optional*) – Dimension of the embedded space.

- **perplexity** (*float, optional*) – Related to the number of nearest neighbors that is used in other manifold learning
- **Larger datasets usually require a larger perplexity. Consider selecting a value between 5 and 50.** (*algorithms.*) –
- **learning_rate** (*float, optional*) – Usually in the range [10.0, 1000.0]. If the cost function gets stuck in a bad
- **minimum** (*local*) –
- **the learning rate may help.** (*increasing*) –
- **metric** (*str, optional*) – The metric to use when calculating distance between instances in a feature array.
- **marker_line_width** (*int, optional*) – Determines the line width of the marker boundary.
- **marker_size** (*int, optional*) – Determines the size of the marker.

Returns `plotly.Figure` representing the transformed data

`evalml.model_understanding.graphs.normalize_confusion_matrix` (*conf_mat, normalize_method='true'*)

Normalizes a confusion matrix.

Parameters

- **conf_mat** (*pd.DataFrame or np.ndarray*) – Confusion matrix to normalize.
- **normalize_method** (*{'true', 'pred', 'all'}*) – Normalization method. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.

Returns normalized version of the input confusion matrix. The column header represents the predicted labels while row header represents the actual labels.

Return type `pd.DataFrame`

`evalml.model_understanding.graphs.partial_dependence` (*pipeline, X, features, percentiles=(0.05, 0.95), grid_resolution=100, kind='average'*)

Calculates one or two-way partial dependence. If a single integer or string is given for features, one-way partial dependence is calculated. If a tuple of two integers or strings is given, two-way partial dependence is calculated with the first feature in the y-axis and second feature in the x-axis.

Parameters

- **pipeline** (*PipelineBase or subclass*) – Fitted pipeline
- **X** (*pd.DataFrame, np.ndarray*) – The input data used to generate a grid of values for feature where partial dependence will be calculated at
- **features** (*int, string, tuple[int or string]*) – The target feature for which to create the partial dependence plot for. If features is an int, it must be the index of the feature to use. If features is a string, it must be a valid column name in X. If features is a tuple of int/strings, it must contain valid column integers/names in X.
- **percentiles** (*tuple[float]*) – The lower and upper percentile used to create the extreme values for the grid. Must be in [0, 1]. Defaults to (0.05, 0.95).

- **grid_resolution** (*int*) – Number of samples of feature(s) for partial dependence plot. If this value is less than the maximum number of categories present in categorical data within X, it will be set to the max number of categories + 1. Defaults to 100.
- **{ 'average' }** (*kind*) – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.
- **'individual'** – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.
- **'both' }** – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.

Returns

When *kind*='average': DataFrame with averaged predictions for all points in the grid averaged over all samples of X and the values used to calculate those predictions.

When *kind*='individual': DataFrame with individual predictions for all points in the grid for each sample of X and the values used to calculate those predictions. If a two-way partial dependence is calculated, then the result is a list of DataFrames with each DataFrame representing one sample's predictions.

When *kind*='both': A tuple consisting of the averaged predictions (in a DataFrame) over all samples of X and the individual predictions (in a list of DataFrames) for each sample of X.

In the one-way case: The dataframe will contain two columns, "feature_values" (grid points at which the partial dependence was calculated) and "partial_dependence" (the partial dependence at that feature value). For classification problems, there will be a third column called "class_label" (the class label for which the partial dependence was calculated). For binary classification, the partial dependence is only calculated for the "positive" class.

In the two-way case: The data frame will contain grid_resolution number of columns and rows where the index and column headers are the sampled values of the first and second features, respectively, used to make the partial dependence contour. The values of the data frame contain the partial dependence data for each feature value pair.

Return type pd.DataFrame, list(pd.DataFrame), or tuple(pd.DataFrame, list(pd.DataFrame))

Raises

- **PartialDependenceError** – if the user provides a tuple of not exactly two features.
- **PartialDependenceError** – if the provided pipeline isn't fitted.
- **PartialDependenceError** – if the provided pipeline is a Baseline pipeline.
- **PartialDependenceError** – if any of the features passed in are completely NaN
- **PartialDependenceError** – if any of the features are low-variance. Defined as having one value occurring more than the upper percentile passed by the user. By default 95%.

`evalml.model_understanding.graphs.precision_recall_curve(y_true, y_pred_proba, pos_label_idx=-1)`

Given labels and binary classifier predicted probabilities, compute and return the data representing a precision-recall curve.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.

- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier, before thresholding has been applied. Note this should be the predicted probability for the “true” label.
- **pos_label_idx** (*int*) – the column index corresponding to the positive class. If predicted probabilities are two-dimensional, this will be used to access the probabilities for the positive class.

Returns

Dictionary containing metrics used to generate a precision-recall plot, with the following keys:

- *precision*: Precision values.
- *recall*: Recall values.
- *thresholds*: Threshold values used to produce the precision and recall.
- *auc_score*: The area under the ROC curve.

Return type list

`evalml.model_understanding.graphs.roc_curve(y_true, y_pred_proba)`

Given labels and classifier predicted probabilities, compute and return the data representing a Receiver Operating Characteristic (ROC) curve. Works with binary or multiclass problems.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True labels.
- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a classifier, before thresholding has been applied.

Returns

A list of dictionaries (with one for each class) is returned. Binary classification problems return a list with one di

Each dictionary contains metrics used to generate an ROC plot with the following keys:

- *fpr_rate*: False positive rate.
- *tpr_rate*: True positive rate.
- *threshold*: Threshold values used to produce each pair of true/false positive rates.
- *auc_score*: The area under the ROC curve.

Return type list(dict)

`evalml.model_understanding.graphs.t_sne(X, n_components=2, perplexity=30.0, learning_rate=200.0, metric='euclidean', **kwargs)`

Get the transformed output after fitting X to the embedded space using t-SNE.

Arguments: X (*np.ndarray*, *pd.DataFrame*): Data to be transformed. Must be numeric.
n_components (*int*, optional): Dimension of the embedded space. perplexity (*float*, optional): Related to the number of nearest neighbors that is used in other manifold learning algorithms. Larger datasets usually require a larger perplexity. Consider selecting a value between 5 and 50. learning_rate (*float*, optional): Usually in the range [10.0, 1000.0]. If the cost function gets stuck in a bad local minimum, increasing the learning rate may help. metric (*str*, optional): The metric to use when calculating distance between instances in a feature array.

Returns *np.ndarray* (n_samples, n_components)

```
evalml.model_understanding.graphs.visualize_decision_tree(estimator,
                                                           max_depth=None,
                                                           rotate=False,
                                                           filled=False,
                                                           filepath=None)
```

Generate an image visualizing the decision tree

Parameters

- **estimator** (*ComponentBase*) – A fitted DecisionTree-based estimator.
- **max_depth** (*int, optional*) – The depth to which the tree should be displayed. If set to None (as by default),
- **is fully generated.** (*tree*) –
- **rotate** (*bool, optional*) – Orient tree left to right rather than top-down.
- **filled** (*bool, optional*) – Paint nodes to indicate majority class for classification, extremity of values for
- **regression** –
- **purity of node for multi-output.** (*or*) –
- **filepath** (*str, optional*) – Path to where the graph should be saved. If set to None (as by default), the graph
- **not be saved.** (*will*) –

Returns DOT object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Source

permutation_importance

Module Contents

Functions

<code>calculate_permutation_importance</code>	Calculates permutation importance for features.
<code>calculate_permutation_importance_one_column</code>	Calculates permutation importance for one column in the original dataframe.

Contents

`evalml.model_understanding.permutation_importance.calculate_permutation_importance` (*pipeline*, *X*, *y*, *objective*, *n_repeats*, *n_jobs*, *random_seed*)

Calculates permutation importance for features.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline.
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance.
- **y** (*pd.Series*) – The target data.
- **objective** (*str*, *ObjectiveBase*) – Objective to score on.
- **n_repeats** (*int*) – Number of times to permute a feature. Defaults to 5.
- **n_jobs** (*int* or *None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For *n_jobs* below -1, $(n_cpus + 1 + n_jobs)$ are used. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Mean feature importance scores over a number of shuffles.

Return type *pd.DataFrame*

`evalml.model_understanding.permutation_importance.calculate_permutation_importance_one_col`

Calculates permutation importance for one column in the original dataframe.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline.
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance.
- **y** (*pd.Series*) – The target data.
- **col_name** (*str*, *int*) – The column in X to calculate permutation importance for.
- **objective** (*str*, *ObjectiveBase*) – Objective to score on.

- **n_repeats** (*int*) – Number of times to permute a feature. Defaults to 5.
- **fast** (*bool*) – Whether to use the fast method of calculating the permutation importance or not. Defaults to True.
- **precomputed_features** (*pd.DataFrame*) – Precomputed features necessary to calculate permutation importance using the fast method. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Mean feature importance scores over a number of shuffles.

Return type float

Package Contents

Functions

<i>binary_objective_vs_threshold</i>	Computes objective score as a function of potential binary classification
<i>calculate_permutation_importance</i>	Calculates permutation importance for features.
<i>calculate_permutation_importance_one_column</i>	Calculates permutation importance for one column in the original dataframe.
<i>confusion_matrix</i>	Confusion matrix for binary and multiclass classification.
<i>explain_predictions</i>	Creates a report summarizing the top contributing features for each data point in the input features.
<i>explain_predictions_best_worst</i>	Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.
<i>get_linear_coefficients</i>	Returns a dataframe showing the features with the greatest predictive power for a linear model.
<i>get_prediction_vs_actual_data</i>	Combines <code>y_true</code> and <code>y_pred</code> into a single dataframe and adds a column for outliers. Used in <i>graph_prediction_vs_actual()</i> .
<i>get_prediction_vs_actual_over_time_data</i>	Get the data needed for the <i>prediction_vs_actual_over_time</i> plot.
<i>graph_binary_objective_vs_threshold</i>	Generates a plot graphing objective score vs. decision thresholds for a fitted binary classification pipeline.
<i>graph_confusion_matrix</i>	Generate and display a confusion matrix plot.
<i>graph_partial_dependence</i>	Create an one-way or two-way partial dependence plot. Passing a single integer or
<i>graph_permutation_importance</i>	Generate a bar graph of the pipeline's permutation importance.
<i>graph_precision_recall_curve</i>	Generate and display a precision-recall plot.
<i>graph_prediction_vs_actual</i>	Generate a scatter plot comparing the true and predicted values. Used for regression plotting
<i>graph_prediction_vs_actual_over_time</i>	Plot the target values and predictions against time on the x-axis.
<i>graph_roc_curve</i>	Generate and display a Receiver Operating Characteristic (ROC) plot for binary and multiclass classification problems.

continues on next page

Table 189 – continued from previous page

<code>graph_t_sne</code>	Plot high dimensional data into lower dimensional space using t-SNE .
<code>normalize_confusion_matrix</code>	Normalizes a confusion matrix.
<code>partial_dependence</code>	Calculates one or two-way partial dependence. If a single integer or
<code>precision_recall_curve</code>	Given labels and binary classifier predicted probabilities, compute and return the data representing a precision-recall curve.
<code>roc_curve</code>	Given labels and classifier predicted probabilities, compute and return the data representing a Receiver Operating Characteristic (ROC) curve. Works with binary or multiclass problems.
<code>t_sne</code>	Get the transformed output after fitting X to the embedded space using t-SNE.

Contents

`evalml.model_understanding.binary_objective_vs_threshold`(*pipeline*, *X*, *y*, *objective*, *steps=100*)

Computes objective score as a function of potential binary classification decision thresholds for a fitted binary classification pipeline.

Parameters

- **pipeline** (*BinaryClassificationPipeline obj*) – Fitted binary classification pipeline
- **X** (*pd.DataFrame*) – The input data used to compute objective score
- **y** (*pd.Series*) – The target labels
- **objective** (*ObjectiveBase obj*, *str*) – Objective used to score
- **steps** (*int*) – Number of intervals to divide and calculate objective score at

Returns DataFrame with thresholds and the corresponding objective score calculated at each threshold

Return type `pd.DataFrame`

`evalml.model_understanding.calculate_permutation_importance`(*pipeline*, *X*, *y*, *objective*, *n_repeats=5*, *n_jobs=None*, *random_seed=0*)

Calculates permutation importance for features.

Parameters

- **pipeline** (*PipelineBase or subclass*) – Fitted pipeline.
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance.
- **y** (*pd.Series*) – The target data.
- **objective** (*str*, *ObjectiveBase*) – Objective to score on.
- **n_repeats** (*int*) – Number of times to permute a feature. Defaults to 5.

- **n_jobs** (*int* or *None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Mean feature importance scores over a number of shuffles.

Return type `pd.DataFrame`

```
evalml.model_understanding.calculate_permutation_importance_one_column(pipeline,
                                                                       X, y,
                                                                       col_name,
                                                                       ob-
                                                                       jec-
                                                                       tive,
                                                                       n_repeats=5,
                                                                       fast=True,
                                                                       pre-
                                                                       com-
                                                                       puted_features=None,
                                                                       ran-
                                                                       dom_seed=0)
```

Calculates permutation importance for one column in the original dataframe.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline.
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance.
- **y** (*pd.Series*) – The target data.
- **col_name** (*str*, *int*) – The column in X to calculate permutation importance for.
- **objective** (*str*, *ObjectiveBase*) – Objective to score on.
- **n_repeats** (*int*) – Number of times to permute a feature. Defaults to 5.
- **fast** (*bool*) – Whether to use the fast method of calculating the permutation importance or not. Defaults to True.
- **precomputed_features** (*pd.DataFrame*) – Precomputed features necessary to calculate permutation importance using the fast method. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Mean feature importance scores over a number of shuffles.

Return type `float`

```
evalml.model_understanding.confusion_matrix(y_true, y_predicted, normal-
                                             ize_method='true')
```

Confusion matrix for binary and multiclass classification.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier.
- **normalize_method** (*{'true', 'pred', 'all', None}*) – Normalization method to use, if not None. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.

Returns Confusion matrix. The column header represents the predicted labels while row header represents the actual labels.

Return type `pd.DataFrame`

```
evalml.model_understanding.explain_predictions(pipeline, input_features, y, indices_to_explain, top_k_features=3, include_shap_values=False, include_expected_value=False, output_format='text', training_data=None, training_target=None)
```

Creates a report summarizing the top contributing features for each data point in the input features.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Dataframe of input data to evaluate the pipeline on.
- **y** (*pd.Series*) – Labels for the input data.
- **indices_to_explain** (*list(int)*) – List of integer indices to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point. Default is 3.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **include_expected_value** (*bool*) – Whether the expected value should be included in the table. Default is False.
- **output_format** (*str*) – Either “text”, “dict”, or “dataframe”. Default is “text”.
- **training_data** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series, np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features to each prediction for each row of
The report will include the feature names, prediction contribution, and SHAP Value (optional).

Raises

- **ValueError** – if `input_features` is empty.
- **ValueError** – if an `output_format` outside of “text”, “dict” or “dataframe” is provided.
- **ValueError** – if the requested index falls outside the `input_feature`’s boundaries.

```
evalml.model_understanding.explain_predictions_best_worst (pipeline,
                                                           input_features, y_true,
                                                           num_to_explain=5,
                                                           top_k_features=3,
                                                           include_shap_values=False,
                                                           metric=None,
                                                           output_format='text',
                                                           callback=None,
                                                           training_data=None,
                                                           training_target=None)
```

Creates a report summarizing the top contributing features for the best and worst points in the dataset as measured by error to true labels.

XGBoost and Stacked Ensemble models, as well as CatBoost multiclass classifiers, are not currently supported.

Parameters

- **pipeline** (*PipelineBase*) – Fitted pipeline whose predictions we want to explain with SHAP.
- **input_features** (*pd.DataFrame*) – Input data to evaluate the pipeline on.
- **y_true** (*pd.Series*) – True labels for the input data.
- **num_to_explain** (*int*) – How many of the best, worst, random data points to explain.
- **top_k_features** (*int*) – How many of the highest/lowest contributing feature to include in the table for each data point.
- **include_shap_values** (*bool*) – Whether SHAP values should be included in the table. Default is False.
- **metric** (*callable*) – The metric used to identify the best and worst points in the dataset. Function must accept the true labels and predicted value or probabilities as the only arguments and lower values must be better. By default, this will be the absolute error for regression problems and cross entropy loss for classification problems.
- **output_format** (*str*) – Either “text” or “dict”. Default is “text”.
- **callback** (*callable*) – Function to be called with incremental updates. Has the following parameters: - *progress_stage*: stage of computation - *time_elapsed*: total time in seconds that has elapsed since start of call
- **training_data** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on. Required and only used for time series pipelines.
- **training_target** (*pd.Series, np.ndarray*) – Targets used to train the pipeline. Required and only used for time series pipelines.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features for the best/worst predictions in the

For each of the best/worst rows of input_features, the predicted values, true labels, metric value, feature names, prediction contribution, and SHAP Value (optional) will be listed.

Raises

- **ValueError** – if input_features does not have more than twice the requested features to explain.
- **ValueError** – if y_true and input_features have mismatched lengths.
- **ValueError** – if an output_format outside of “text”, “dict” or “dataframe” is provided.

`evalml.model_understanding.get_linear_coefficients` (*estimator*, *features=None*)

Returns a dataframe showing the features with the greatest predictive power for a linear model.

Parameters

- **estimator** (*Estimator*) – Fitted linear model family estimator.
- **features** (*list[str]*) – List of feature names associated with the underlying data.

Returns Displaying the features by importance.

Return type `pd.DataFrame`

`evalml.model_understanding.get_prediction_vs_actual_data` (*y_true*, *y_pred*, *outlier_threshold=None*)

Combines *y_true* and *y_pred* into a single dataframe and adds a column for outliers. Used in `graph_prediction_vs_actual()`.

Parameters

- **y_true** (*pd.Series*, or *np.ndarray*) – The real target values of the data
- **y_pred** (*pd.Series*, or *np.ndarray*) – The predicted values outputted by the regression model.
- **outlier_threshold** (*int*, *float*) – A positive threshold for what is considered an outlier value. This value is compared to the absolute difference between each value of *y_true* and *y_pred*. Values within this threshold will be blue, otherwise they will be yellow. Defaults to None

Returns

- *prediction*: Predicted values from regression model.
- *actual*: Real target values.
- *outlier*: Colors indicating which values are in the threshold for what is considered an outlier value.

Return type `pd.DataFrame` with the following columns

`evalml.model_understanding.get_prediction_vs_actual_over_time_data` (*pipeline*, *X*, *y*, *X_train*, *y_train*, *dates*)

Get the data needed for the `prediction_vs_actual_over_time` plot.

Parameters

- **pipeline** (*TimeSeriesRegressionPipeline*) – Fitted time series regression pipeline.
- **X** (*pd.DataFrame*) – Features used to generate new predictions.
- **y** (*pd.Series*) – Target values to compare predictions against.
- **X_train** (*pd.DataFrame*) – Data the pipeline was trained on.
- **y_train** (*pd.Series*) – Target values for training data.
- **dates** (*pd.Series*) – Dates corresponding to target values and predictions.

Returns `pd.DataFrame`

```
evalml.model_understanding.graph_binary_objective_vs_threshold(pipeline, X,
                                                             y, objective,
                                                             steps=100)
```

Generates a plot graphing objective score vs. decision thresholds for a fitted binary classification pipeline.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*) – The input data used to score and compute scores
- **y** (*pd.Series*) – The target labels
- **objective** (*ObjectiveBase obj, str*) – Objective used to score, shown on the y-axis of the graph
- **steps** (*int*) – Number of intervals to divide and calculate objective score at

Returns *plotly.Figure* representing the objective score vs. threshold graph generated

```
evalml.model_understanding.graph_confusion_matrix(y_true, y_pred, normalize_method='true',
                                                  title_addition=None)
```

Generate and display a confusion matrix plot.

If *normalize_method* is set, hover text will show raw count, otherwise hover text will show count normalized with method 'true'.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier.
- **normalize_method** (*{'true', 'pred', 'all', None}*) – Normalization method to use, if not None. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.
- **title_addition** (*str* or *None*) – if not None, append to plot title. Defaults to None.

Returns *plotly.Figure* representing the confusion matrix plot generated

```
evalml.model_understanding.graph_partial_dependence(pipeline, X, features,
                                                    class_label=None,
                                                    grid_resolution=100,
                                                    kind='average')
```

Create an one-way or two-way partial dependence plot. Passing a single integer or string as features will create a one-way partial dependence plot with the feature values plotted against the partial dependence. Passing features a tuple of int/strings will create a two-way partial dependence plot with a contour of feature[0] in the y-axis, feature[1] in the x-axis and the partial dependence in the z-axis.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline
- **X** (*pd.DataFrame, np.ndarray*) – The input data used to generate a grid of values for feature where partial dependence will be calculated at
- **features** (*int, string, tuple[int or string]*) – The target feature for which to create the partial dependence plot for. If features is an int, it must be the index of the feature to use. If features is a string, it must be a valid column name in X. If features is a tuple of strings, it must contain valid column int/names in X.

- **class_label** (*string, optional*) – Name of class to plot for multiclass problems. If None, will plot the partial dependence for each class. This argument does not change behavior for regression or binary classification pipelines. For binary classification, the partial dependence for the positive label will always be displayed. Defaults to None.
- **grid_resolution** (*int*) – Number of samples of feature(s) for partial dependence plot
- **{ 'average' }** (*kind*) – Type of partial dependence to plot. ‘average’ creates a regular partial dependence (PD) graph, ‘individual’ creates an individual conditional expectation (ICE) plot, and ‘both’ creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.
- **'individual'** – Type of partial dependence to plot. ‘average’ creates a regular partial dependence (PD) graph, ‘individual’ creates an individual conditional expectation (ICE) plot, and ‘both’ creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.
- **'both'** } – Type of partial dependence to plot. ‘average’ creates a regular partial dependence (PD) graph, ‘individual’ creates an individual conditional expectation (ICE) plot, and ‘both’ creates a single-figure PD and ICE plot. ICE plots can only be shown for one-way partial dependence plots.

Returns figure object containing the partial dependence data for plotting

Return type `plotly.graph_objects.Figure`

Raises

- **PartialDependenceError** – if a graph is requested for a class name that isn’t present in the pipeline.
- **PartialDependenceError** – if an ICE plot is requested for a two-way partial dependence.

`evalml.model_understanding.graph_permutation_importance` (*pipeline, X, y, objective, importance_threshold=0*)

Generate a bar graph of the pipeline’s permutation importance.

Parameters

- **pipeline** (*PipelineBase or subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*) – The input data used to score and compute permutation importance
- **y** (*pd.Series*) – The target data
- **objective** (*str, ObjectiveBase*) – Objective to score on
- **importance_threshold** (*float, optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their respective permutation importance.

`evalml.model_understanding.graph_precision_recall_curve` (*y_true, y_pred_proba, title_addition=None*)

Generate and display a precision-recall plot.

Parameters

- **y_true** (*pd.Series or np.ndarray*) – True binary labels.
- **y_pred_proba** (*pd.Series or np.ndarray*) – Predictions from a binary classifier, before thresholding has been applied. Note this should be the predicted probability for the “true” label.

- **title_addition** (*str* or *None*) – If not *None*, append to plot title. Default *None*.

Returns `plotly.Figure` representing the precision-recall plot generated

`evalml.model_understanding.graph_prediction_vs_actual` (*y_true*, *y_pred*, *outlier_threshold=None*)

Generate a scatter plot comparing the true and predicted values. Used for regression plotting

Parameters

- **y_true** (*pd.Series*) – The real target values of the data
- **y_pred** (*pd.Series*) – The predicted values outputted by the regression model.
- **outlier_threshold** (*int*, *float*) – A positive threshold for what is considered an outlier value. This value is compared to the absolute difference between each value of *y_true* and *y_pred*. Values within this threshold will be blue, otherwise they will be yellow. Defaults to *None*

Returns `plotly.Figure` representing the predicted vs. actual values graph

`evalml.model_understanding.graph_prediction_vs_actual_over_time` (*pipeline*, *X*, *y*, *X_train*, *y_train*, *dates*)

Plot the target values and predictions against time on the x-axis.

Parameters

- **pipeline** (*TimeSeriesRegressionPipeline*) – Fitted time series regression pipeline.
- **X** (*pd.DataFrame*) – Features used to generate new predictions.
- **y** (*pd.Series*) – Target values to compare predictions against.
- **X_train** (*pd.DataFrame*) – Data the pipeline was trained on.
- **y_train** (*pd.Series*) – Target values for training data.
- **dates** (*pd.Series*) – Dates corresponding to target values and predictions.

Returns Showing the prediction vs actual over time.

Return type `plotly.Figure`

`evalml.model_understanding.graph_roc_curve` (*y_true*, *y_pred_proba*, *custom_class_names=None*, *title_addition=None*)

Generate and display a Receiver Operating Characteristic (ROC) plot for binary and multiclass classification problems.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True labels.
- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a classifier, before thresholding has been applied. Note this should be a one dimensional array with the predicted probability for the “true” label in the binary case.
- **custom_class_labels** (*list* or *None*) – If not *None*, custom labels for classes. Default *None*.
- **title_addition** (*str* or *None*) – if not *None*, append to plot title. Default *None*.

Returns `plotly.Figure` representing the ROC plot generated

```
evalml.model_understanding.graph_t_sne(X, n_components=2, perplexity=30.0,  
                                       learning_rate=200.0, metric='euclidean',  
                                       marker_line_width=2, marker_size=7, **kwargs)
```

Plot high dimensional data into lower dimensional space using t-SNE .

Parameters

- **X** (*np.ndarray*, *pd.DataFrame*) – Data to be transformed. Must be numeric.
- **n_components** (*int*, *optional*) – Dimension of the embedded space.
- **perplexity** (*float*, *optional*) – Related to the number of nearest neighbors that is used in other manifold learning
- **Larger datasets usually require a larger perplexity. Consider selecting a value between 5 and 50.** (*algorithms.*)
–
- **learning_rate** (*float*, *optional*) – Usually in the range [10.0, 1000.0]. If the cost function gets stuck in a bad
- **minimum** (*local*) –
- **the learning rate may help.** (*increasing*) –
- **metric** (*str*, *optional*) – The metric to use when calculating distance between instances in a feature array.
- **marker_line_width** (*int*, *optional*) – Determines the line width of the marker boundary.
- **marker_size** (*int*, *optional*) – Determines the size of the marker.

Returns `plotly.Figure` representing the transformed data

```
evalml.model_understanding.normalize_confusion_matrix(conf_mat, normal-  
                                                    ize_method='true')
```

Normalizes a confusion matrix.

Parameters

- **conf_mat** (*pd.DataFrame* or *np.ndarray*) – Confusion matrix to normalize.
- **normalize_method** (*{'true', 'pred', 'all'}*) – Normalization method. Supported options are: 'true' to normalize by row, 'pred' to normalize by column, or 'all' to normalize by all values. Defaults to 'true'.

Returns normalized version of the input confusion matrix. The column header represents the predicted labels while row header represents the actual labels.

Return type `pd.DataFrame`

```
evalml.model_understanding.partial_dependence(pipeline, X, features, percentiles=(0.05,  
                                                                                   0.95),  
                                                                                   grid_resolution=100,  
                                                                                   kind='average')
```

Calculates one or two-way partial dependence. If a single integer or string is given for features, one-way partial dependence is calculated. If a tuple of two integers or strings is given, two-way partial dependence is calculated with the first feature in the y-axis and second feature in the x-axis.

Parameters

- **pipeline** (*PipelineBase* or *subclass*) – Fitted pipeline
- **X** (*pd.DataFrame*, *np.ndarray*) – The input data used to generate a grid of values for feature where partial dependence will be calculated at

- **features** (*int, string, tuple[int or string]*) – The target feature for which to create the partial dependence plot for. If features is an int, it must be the index of the feature to use. If features is a string, it must be a valid column name in X. If features is a tuple of int/strings, it must contain valid column integers/names in X.
- **percentiles** (*tuple[float]*) – The lower and upper percentile used to create the extreme values for the grid. Must be in [0, 1]. Defaults to (0.05, 0.95).
- **grid_resolution** (*int*) – Number of samples of feature(s) for partial dependence plot. If this value is less than the maximum number of categories present in categorical data within X, it will be set to the max number of categories + 1. Defaults to 100.
- **{ 'average' (kind) }** – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.
- **'individual'** – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.
- **'both' }** – The type of predictions to return. 'individual' will return the predictions for all of the points in the grid for each sample in X. 'average' will return the predictions for all of the points in the grid but averaged over all of the samples in X.

Returns

When *kind='average'*: DataFrame with averaged predictions for all points in the grid averaged over all samples of X and the values used to calculate those predictions.

When *kind='individual'*: DataFrame with individual predictions for all points in the grid for each sample of X and the values used to calculate those predictions. If a two-way partial dependence is calculated, then the result is a list of DataFrames with each DataFrame representing one sample's predictions.

When *kind='both'*: A tuple consisting of the averaged predictions (in a DataFrame) over all samples of X and the individual predictions (in a list of DataFrames) for each sample of X.

In the one-way case: The dataframe will contain two columns, "feature_values" (grid points at which the partial dependence was calculated) and "partial_dependence" (the partial dependence at that feature value). For classification problems, there will be a third column called "class_label" (the class label for which the partial dependence was calculated). For binary classification, the partial dependence is only calculated for the "positive" class.

In the two-way case: The data frame will contain grid_resolution number of columns and rows where the index and column headers are the sampled values of the first and second features, respectively, used to make the partial dependence contour. The values of the data frame contain the partial dependence data for each feature value pair.

Return type pd.DataFrame, list(pd.DataFrame), or tuple(pd.DataFrame, list(pd.DataFrame))

Raises

- **PartialDependenceError** – if the user provides a tuple of not exactly two features.
- **PartialDependenceError** – if the provided pipeline isn't fitted.
- **PartialDependenceError** – if the provided pipeline is a Baseline pipeline.
- **PartialDependenceError** – if any of the features passed in are completely NaN
- **PartialDependenceError** – if any of the features are low-variance. Defined as having one value occurring more than the upper percentile passed by the user. By default 95%.

`evalml.model_understanding.precision_recall_curve` (*y_true*, *y_pred_proba*,
pos_label_idx=-1)

Given labels and binary classifier predicted probabilities, compute and return the data representing a precision-recall curve.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True binary labels.
- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a binary classifier, before thresholding has been applied. Note this should be the predicted probability for the “true” label.
- **pos_label_idx** (*int*) – the column index corresponding to the positive class. If predicted probabilities are two-dimensional, this will be used to access the probabilities for the positive class.

Returns

Dictionary containing metrics used to generate a precision-recall plot, with the following keys:

- *precision*: Precision values.
- *recall*: Recall values.
- *thresholds*: Threshold values used to produce the precision and recall.
- *auc_score*: The area under the ROC curve.

Return type

 list

`evalml.model_understanding.roc_curve` (*y_true*, *y_pred_proba*)

Given labels and classifier predicted probabilities, compute and return the data representing a Receiver Operating Characteristic (ROC) curve. Works with binary or multiclass problems.

Parameters

- **y_true** (*pd.Series* or *np.ndarray*) – True labels.
- **y_pred_proba** (*pd.Series* or *np.ndarray*) – Predictions from a classifier, before thresholding has been applied.

Returns

A list of dictionaries (with one for each class) is returned. Binary classification problems return a list with one di

Each dictionary contains metrics used to generate an ROC plot with the following keys:

- *fpr_rate*: False positive rate.
- *tpr_rate*: True positive rate.
- *threshold*: Threshold values used to produce each pair of true/false positive rates.
- *auc_score*: The area under the ROC curve.

Return type

 list(dict)

`evalml.model_understanding.t_sne` (*X*, *n_components=2*, *perplexity=30.0*, *learning_rate=200.0*,
metric='euclidean', ***kwargs*)

Get the transformed output after fitting *X* to the embedded space using t-SNE.

Arguments: *X* (*np.ndarray*, *pd.DataFrame*): Data to be transformed. Must be numeric.
n_components (*int*, optional): Dimension of the embedded space. *perplexity* (*float*, optional):
Related to the number of nearest neighbors that is used in other manifold learning algorithms.

Larger datasets usually require a larger perplexity. Consider selecting a value between 5 and 50. `learning_rate` (float, optional): Usually in the range [10.0, 1000.0]. If the cost function gets stuck in a bad local minimum, increasing the learning rate may help. `metric` (str, optional): The metric to use when calculating distance between instances in a feature array.

Returns np.ndarray (n_samples, n_components)

Objectives

Submodules

binary_classification_objective

Module Contents

Classes Summary

<i>BinaryClassificationObjective</i>	Base class for all binary classification objectives.
--------------------------------------	--

Contents

class evalml.objectives.binary_classification_objective.**BinaryClassificationObjective**
Base class for all binary classification objectives.

Attributes

problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
----------------------	--

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<i>greater_is_better</i>	Returns a boolean determining if a greater score indicates better model performance.
<i>is_bounded_like_percentage</i>	Returns whether this objective is bounded between 0 and 1, inclusive.
<i>is_defined_for_problem_type</i>	
<i>name</i>	Returns a name describing the objective.

continues on next page

Table 191 – continued from previous page

<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>perfect_score</i>	Returns the score obtained by evaluating this objective on a perfect model.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>score_needs_proba</i>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

property expected_range (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property greater_is_better (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property is_bounded_like_percentage (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod is_defined_for_problem_type (*cls, problem_type*)

property name (*cls*)

Returns a name describing the objective.

abstract classmethod objective_function (*cls, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

property perfect_score (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame or np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

property `score_needs_proba` (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

cost_benefit_matrix**Module Contents****Classes Summary**

CostBenefitMatrix

Score using a cost-benefit matrix. Scores quantify the benefits of a given value, so greater numeric

Contents

```
class evalml.objectives.cost_benefit_matrix.CostBenefitMatrix(true_positive,  
                                                             true_negative,  
                                                             false_positive,  
                                                             false_negative)
```

Score using a cost-benefit matrix. Scores quantify the benefits of a given value, so greater numeric scores represents a better score. Costs and scores can be negative, indicating that a value is not beneficial. For example, in the case of monetary profit, a negative cost and/or score represents loss of cash flow.

Parameters

- **true_positive** (*float*) – Cost associated with true positive predictions
- **true_negative** (*float*) – Cost associated with true negative predictions
- **false_positive** (*float*) – Cost associated with false positive predictions
- **false_negative** (*float*) – Cost associated with false negative predictions

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Cost Benefit Matrix
per-fect_score	None
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculates cost-benefit of the using the predicted and true values.
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series, np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float, optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Calculates cost-benefit of the using the predicted and true values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted labels
- **y_true** (*pd.Series*) – True labels
- **X** (*pd.DataFrame*) – Ignored.
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns Cost-benefit matrix score

Return type float

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame or np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

fraud_cost

Module Contents

Classes Summary

<i>FraudCost</i>	Score the percentage of money lost of the total transaction amount process due to fraud.
------------------	--

Contents

class evalml.objectives.fraud_cost.**FraudCost** (*retry_percentage=0.5*, *interchange_fee=0.02*, *fraud_payout_percentage=1.0*, *amount_col='amount'*) *in-*

Score the percentage of money lost of the total transaction amount process due to fraud.

Parameters

- **retry_percentage** (*float*) – What percentage of customers that will retry a transaction if it is declined. Between 0 and 1. Defaults to 0.5.
- **interchange_fee** (*float*) – How much of each successful transaction you pay. Between 0 and 1. Defaults to 0.02.
- **fraud_payout_percentage** (*float*) – Percentage of fraud you will not be able to collect. Between 0 and 1. Defaults to 1.0.
- **amount_col** (*str*) – Name of column in data that contains the amount. Defaults to “amount”.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	Fraud Cost
perfect_score	0.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Calculate amount lost to fraud per transaction given predictions, true values, and dataframe with transaction amount.
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X, sample_weight=None*)

Calculate amount lost to fraud per transaction given predictions, true values, and dataframe with transaction amount.

Parameters

- **y_predicted** (*pd.Series*) – Predicted fraud labels
- **y_true** (*pd.Series*) – True fraud labels
- **x** (*pd.DataFrame*) – Data with transaction amounts
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns Amount lost to fraud per transaction

Return type float

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **x** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **x** (*pd.DataFrame or np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series, or pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

lead_scoring

Module Contents

Classes Summary

<i>LeadScoring</i>	Lead scoring.
--------------------	---------------

Contents

class evalml.objectives.lead_scoring.**LeadScoring**(*true_positives=1*, *false_positives=-1*)

Lead scoring.

Parameters

- **true_positives** (*int*) – Reward for a true positive. Defaults to 1.
- **false_positives** (*int*) – Cost for a false positive. Should be negative. Defaults to -1.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Lead Scoring
per-fect_score	None
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculate the profit per lead.
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self, ypred_proba, threshold=0.5, X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series, np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float, optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Calculate the profit per lead.

Parameters

- **y_predicted** (*pd.Series*) – Predicted labels
- **y_true** (*pd.Series*) – True labels
- **X** (*pd.DataFrame*) – Ignored.
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns Profit per lead

Return type `float`

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.

- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

multiclass_classification_objective

Module Contents

Classes Summary

<i>MulticlassClassificationObjective</i>	Base class for all multiclass classification objectives.
--	--

Contents

class evalml.objectives.multiclass_classification_objective.**MulticlassClassificationObjective**

Base class for all multiclass classification objectives.

Attributes

problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
----------------------	--

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>expected_range</code>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<code>greater_is_better</code>	Returns a boolean determining if a greater score indicates better model performance.
<code>is_bounded_like_percentage</code>	Returns whether this objective is bounded between 0 and 1, inclusive.
<code>is_defined_for_problem_type</code>	
<code>name</code>	Returns a name describing the objective.
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>perfect_score</code>	Returns the score obtained by evaluating this objective on a perfect model.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod `objective_function` (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property `perfect_score` (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

property `positive_only` (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property `score_needs_proba` (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

objective_base

Module Contents

Classes Summary

<i>ObjectiveBase</i>	Base class for all objectives.
----------------------	--------------------------------

Contents

class evalml.objectives.objective_base.**ObjectiveBase**

Base class for all objectives.

Attributes

problem_types	None
----------------------	------

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<i>greater_is_better</i>	Returns a boolean determining if a greater score indicates better model performance.
<i>is_bounded_like_percentage</i>	Returns whether this objective is bounded between 0 and 1, inclusive.
<i>is_defined_for_problem_type</i>	
<i>name</i>	Returns a name describing the objective.
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>perfect_score</i>	Returns the score obtained by evaluating this objective on a perfect model.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>score_needs_proba</i>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod `objective_function` (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property `perfect_score` (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

property `positive_only` (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score

- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property score_needs_proba (*cls*)

Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

regression_objective

Module Contents

Classes Summary

RegressionObjective

Base class for all regression objectives.

Contents

class evalml.objectives.regression_objective.**RegressionObjective**

Base class for all regression objectives.

Attributes

problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
----------------------	--

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<i>greater_is_better</i>	Returns a boolean determining if a greater score indicates better model performance.
<i>is_bounded_like_percentage</i>	Returns whether this objective is bounded between 0 and 1, inclusive.

continues on next page

Table 203 – continued from previous page

<i>is_defined_for_problem_type</i>	
<i>name</i>	Returns a name describing the objective.
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>perfect_score</i>	Returns the score obtained by evaluating this objective on a perfect model.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>score_needs_proba</i>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod `objective_function` (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (`pd.Series`): Predicted values of length `[n_samples]` `y_true` (`pd.Series`): Actual class labels of length `[n_samples]` `X` (`pd.DataFrame` or `np.ndarray`): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (`pd.DataFrame` or `np.ndarray`): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property `perfect_score` (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

property `positive_only` (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **`y_predicted`** (*pd.Series*) – Predicted values of length `[n_samples]`
- **`y_true`** (*pd.Series*) – Actual class labels of length `[n_samples]`
- **`X`** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape `[n_samples, n_features]` necessary to calculate score
- **`sample_weight`** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property `score_needs_proba` (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **`y_predicted`** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length `[n_samples]`
- **`y_true`** (*pd.Series*) – Actual class labels of length `[n_samples]`

Returns None

sensitivity_low_alert

Module Contents

Classes Summary

SensitivityLowAlert

Base class for all binary classification objectives.

Attributes Summary

logger

Contents

`evalml.objectives.sensitivity_low_alert.logger`

class `evalml.objectives.sensitivity_low_alert.SensitivityLowAlert` (*alert_rate=0.01*)
 Base class for all binary classification objectives.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Sensitivity at Low Alert Rates
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Determine if an observation is high risk given an alert rate
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculate sensitivity across all predictions, using the top alert_rate percent of observations as the predicted positive class
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.

- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property can_optimize_threshold (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, ***kwargs*)

Determine if an observation is high risk given an alert rate

Parameters *ypred_proba* (*pd.Series*) – Predicted probabilities

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, ***kwargs*)

Calculate sensitivity across all predictions, using the top alert_rate percent of observations as the predicted positive class

Parameters

- **y_true** (*pd.Series*) – True labels
- **y_predicted** (*pd.Series*) – Predicted labels based on alert_rate

Returns sensitivity using the observations with the top scores as the predicted positive class

Return type float

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score

- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

standard_metrics

Module Contents

Classes Summary

<i>AccuracyBinary</i>	Accuracy score for binary classification.
<i>AccuracyMulticlass</i>	Accuracy score for multiclass classification.
<i>AUC</i>	AUC score for binary classification.
<i>AUCMacro</i>	AUC score for multiclass classification using macro averaging.
<i>AUCMicro</i>	AUC score for multiclass classification using micro averaging.
<i>AUCWeighted</i>	AUC Score for multiclass classification using weighted averaging.
<i>BalancedAccuracyBinary</i>	Balanced accuracy score for binary classification.
<i>BalancedAccuracyMulticlass</i>	Balanced accuracy score for multiclass classification.
<i>ExpVariance</i>	Explained variance score for regression.
<i>F1</i>	F1 score for binary classification.
<i>F1Macro</i>	F1 score for multiclass classification using macro averaging.
<i>F1Micro</i>	F1 score for multiclass classification using micro averaging.
<i>F1Weighted</i>	F1 score for multiclass classification using weighted averaging.
<i>Gini</i>	Gini coefficient for binary classification.
<i>LogLossBinary</i>	Log Loss for binary classification.
<i>LogLossMulticlass</i>	Log Loss for multiclass classification.
<i>MAE</i>	Mean absolute error for regression.
<i>MAPE</i>	Mean absolute percentage error for time series regression. Scaled by 100 to return a percentage.
<i>MaxError</i>	Maximum residual error for regression.
<i>MCCBinary</i>	Matthews correlation coefficient for binary classification.
<i>MCCMulticlass</i>	Matthews correlation coefficient for multiclass classification.

continues on next page

Table 207 – continued from previous page

<i>MeanSquaredLogError</i>	Mean squared log error for regression.
<i>MedianAE</i>	Median absolute error for regression.
<i>MSE</i>	Mean squared error for regression.
<i>Precision</i>	Precision score for binary classification.
<i>PrecisionMacro</i>	Precision score for multiclass classification using macro averaging.
<i>PrecisionMicro</i>	Precision score for multiclass classification using micro averaging.
<i>PrecisionWeighted</i>	Precision score for multiclass classification using weighted averaging.
<i>R2</i>	Coefficient of determination for regression.
<i>Recall</i>	Recall score for binary classification.
<i>RecallMacro</i>	Recall score for multiclass classification using macro averaging.
<i>RecallMicro</i>	Recall score for multiclass classification using micro averaging.
<i>RecallWeighted</i>	Recall score for multiclass classification using weighted averaging.
<i>RootMeanSquaredError</i>	Root mean squared error for regression.
<i>RootMeanSquaredLogError</i>	Root mean squared log error for regression.

Contents

class evalml.objectives.standard_metrics.**AccuracyBinary**

Accuracy score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Accuracy Binary
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	

continues on next page

Table 208 – continued from previous page

<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**AccuracyMulticlass**

Accuracy score for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Accuracy Multiclass
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**AUC**

AUC score for binary classification.

Attributes

expected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self, ypred_proba, threshold=0.5, X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series, np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float, optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (*pd.Series*): Predicted values of length `[n_samples]` `y_true` (*pd.Series*): Actual class labels of length `[n_samples]` `X` (*pd.DataFrame* or *np.ndarray*): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**AUCMacro**

AUC score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 211 – continued from previous page

<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.standard_metrics.**AUCMicro**

AUC score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Micro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**AUCWeighted**

AUC Score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Weighted
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**BalancedAccuracyBinary**

Balanced accuracy score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Balanced Accuracy Binary
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self, ypred_proba, threshold=0.5, X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series, np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float, optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (*pd.Series*): Predicted values of length `[n_samples]` `y_true` (*pd.Series*): Actual class labels of length `[n_samples]` `X` (*pd.DataFrame* or *np.ndarray*): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**BalancedAccuracyMulticlass**

Balanced accuracy score for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Balanced Accuracy Multiclass
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 215 – continued from previous page

<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod calculate_percent_difference (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.standard_metrics.**ExpVariance**

Explained variance score for regression.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	ExpVariance
per-fect_score	1.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**F1**

F1 score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (`pd.Series`): Predicted values of length `[n_samples]` `y_true` (`pd.Series`): Actual class labels of length `[n_samples]` `X` (`pd.DataFrame` or `np.ndarray`): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (`pd.DataFrame` or `np.ndarray`): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (`pd.Series`) – The classifier’s predicted probabilities
- **y_true** (`pd.Series`) – The ground truth for the predictions.
- **X** (`pd.DataFrame`, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (`pd.Series`) – Predicted values of length `[n_samples]`
- **y_true** (`pd.Series`) – Actual class labels of length `[n_samples]`
- **X** (`pd.DataFrame` or `np.ndarray`) – Extra data of shape `[n_samples, n_features]` necessary to calculate score
- **sample_weight** (`pd.DataFrame` or `np.ndarray`) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (`pd.Series`, or `pd.DataFrame`) – Predicted values of length `[n_samples]`
- **y_true** (`pd.Series`) – Actual class labels of length `[n_samples]`

Returns None

class `evalml.objectives.standard_metrics.F1Macro`

F1 score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**F1Micro**

F1 score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Micro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 219 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns `score`

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**F1Weighted**

F1 score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Weighted
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.standard_metrics.**Gini**

Gini coefficient for binary classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Gini
per-fect_score	1.0
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.standard_metrics.**LogLossBinary**

Log Loss for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	False
is_bounded_like_percentage	False
name	Log Loss Binary
perfect_score	0.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**LogLossMulticlass**

Log Loss for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	False
is_bounded_like_percentage	True
name	Log Loss Multiclass
perfect_score	0.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- ***y_predicted*** (*pd.Series*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]
- ***X*** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- ***sample_weight*** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- ***y_predicted*** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.standard_metrics.MAE`

Mean absolute error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	MAE
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**MAPE**

Mean absolute percentage error for time series regression. Scaled by 100 to return a percentage.

Only valid for nonzero inputs. Otherwise, will throw a ValueError

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	Mean Absolute Percentage Error
per-fect_score	0.0
prob-lem_types	[ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 225 – continued from previous page

<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**MaxError**
Maximum residual error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MaxError
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)
Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.MCCBinary

Matthews correlation coefficient for binary classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	MCC Binary
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.standard_metrics.MCCMulticlass`

Matthews correlation coefficient for multiclass classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	MCC Multiclass
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame or np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series, or pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**MeanSquaredLogError**

Mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Mean Squared Log Error
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 229 – continued from previous page

<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.standard_metrics.**MedianAE**

Median absolute error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MedianAE
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.standard_metrics.MSE`

Mean squared error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MSE
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**Precision**

Precision score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**PrecisionMacro**

Precision score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Macro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 233 – continued from previous page

<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod calculate_percent_difference (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.standard_metrics.**PrecisionMicro**

Precision score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Micro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**PrecisionWeighted**

Precision score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Weighted
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**R2**

Coefficient of determination for regression.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	R2
perfect_score	1
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.

- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod is_defined_for_problem_type (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.standard_metrics.**Recall**

Recall score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**RecallMacro**

Recall score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (`pd.Series`): Predicted values of length `[n_samples]` `y_true` (`pd.Series`): Actual class labels of length `[n_samples]` `X` (`pd.DataFrame` or `np.ndarray`): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (`pd.DataFrame` or `np.ndarray`): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **`y_predicted`** (`pd.Series`) – Predicted values of length `[n_samples]`
- **`y_true`** (`pd.Series`) – Actual class labels of length `[n_samples]`
- **`X`** (`pd.DataFrame` or `np.ndarray`) – Extra data of shape `[n_samples, n_features]` necessary to calculate score
- **`sample_weight`** (`pd.DataFrame` or `np.ndarray`) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **`y_predicted`** (`pd.Series`, or `pd.DataFrame`) – Predicted values of length `[n_samples]`
- **`y_true`** (`pd.Series`) – Actual class labels of length `[n_samples]`

Returns None

class `evalml.objectives.standard_metrics.RecallMicro`

Recall score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Micro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**RecallWeighted**

Recall score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Weighted
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 240 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.standard_metrics.**RootMeanSquaredError**

Root mean squared error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Root Mean Squared Error
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.standard_metrics.**RootMeanSquaredLogError**

Root mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Root Mean Squared Log Error
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

time_series_regression_objective

Module Contents

Classes Summary

<i>TimeSeriesRegressionObjective</i>	Base class for all time series regression objectives.
--------------------------------------	---

Contents

class evalml.objectives.time_series_regression_objective.**TimeSeriesRegressionObjective**

Base class for all time series regression objectives.

Attributes

problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]
----------------------	---------------------------------------

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<i>greater_is_better</i>	Returns a boolean determining if a greater score indicates better model performance.
<i>is_bounded_like_percentage</i>	Returns whether this objective is bounded between 0 and 1, inclusive.
<i>is_defined_for_problem_type</i>	
<i>name</i>	Returns a name describing the objective.

continues on next page

Table 244 – continued from previous page

<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>perfect_score</code>	Returns the score obtained by evaluating this objective on a perfect model.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the <code>score()</code> method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod `objective_function` (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of

shape [n_samples, n_features] necessary to calculate score sample_weight (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property perfect_score (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property score_needs_proba (*cls*)

Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

utils

Module Contents

Functions

<i>get_all_objective_names</i>	Get a list of the names of all objectives.
<i>get_core_objective_names</i>	Get a list of all valid core objectives.
<i>get_core_objectives</i>	Returns all core objective instances associated with the given problem type.
<i>get_non_core_objectives</i>	Get non-core objective classes.

continues on next page

Table 245 – continued from previous page

<code>get_objective</code>	Returns the Objective class corresponding to a given objective name.
----------------------------	--

Contents

`evalml.objectives.utils.get_all_objective_names()`

Get a list of the names of all objectives.

Returns Objective names

Return type list(str)

`evalml.objectives.utils.get_core_objective_names()`

Get a list of all valid core objectives.

Returns Objective names.

Return type list[str]

`evalml.objectives.utils.get_core_objectives(problem_type)`

Returns all core objective instances associated with the given problem type.

Core objectives are designed to work out-of-the-box for any dataset.

Parameters `problem_type` (str/*ProblemTypes*) – Type of problem

Returns List of ObjectiveBase instances

`evalml.objectives.utils.get_non_core_objectives()`

Get non-core objective classes.

Non-core objectives are objectives that are domain-specific. Users typically need to configure these objectives before using them in AutoMLSearch.

Returns List of ObjectiveBase classes

`evalml.objectives.utils.get_objective(objective, return_instance=False, **kwargs)`

Returns the Objective class corresponding to a given objective name.

Parameters

- **objective** (str or *ObjectiveBase*) – Name or instance of the objective class.
- **return_instance** (bool) – Whether to return an instance of the objective. This only applies if objective is of type str. Note that the instance will be initialized with default arguments.
- **kwargs** (*Any*) – Any keyword arguments to pass into the objective. Only used when `return_instance=True`.

Returns ObjectiveBase if the parameter objective is of type ObjectiveBase. If objective is instead a valid objective name, function will return the class corresponding to that name. If `return_instance` is True, an instance of that objective will be returned.

Package Contents

Classes Summary

<i>AccuracyBinary</i>	Accuracy score for binary classification.
<i>AccuracyMulticlass</i>	Accuracy score for multiclass classification.
<i>AUC</i>	AUC score for binary classification.
<i>AUCMacro</i>	AUC score for multiclass classification using macro averaging.
<i>AUCMicro</i>	AUC score for multiclass classification using micro averaging.
<i>AUCWeighted</i>	AUC Score for multiclass classification using weighted averaging.
<i>BalancedAccuracyBinary</i>	Balanced accuracy score for binary classification.
<i>BalancedAccuracyMulticlass</i>	Balanced accuracy score for multiclass classification.
<i>BinaryClassificationObjective</i>	Base class for all binary classification objectives.
<i>CostBenefitMatrix</i>	Score using a cost-benefit matrix. Scores quantify the benefits of a given value, so greater numeric
<i>ExpVariance</i>	Explained variance score for regression.
<i>F1</i>	F1 score for binary classification.
<i>F1Macro</i>	F1 score for multiclass classification using macro averaging.
<i>F1Micro</i>	F1 score for multiclass classification using micro averaging.
<i>F1Weighted</i>	F1 score for multiclass classification using weighted averaging.
<i>FraudCost</i>	Score the percentage of money lost of the total transaction amount process due to fraud.
<i>Gini</i>	Gini coefficient for binary classification.
<i>LeadScoring</i>	Lead scoring.
<i>LogLossBinary</i>	Log Loss for binary classification.
<i>LogLossMulticlass</i>	Log Loss for multiclass classification.
<i>MAE</i>	Mean absolute error for regression.
<i>MAPE</i>	Mean absolute percentage error for time series regression. Scaled by 100 to return a percentage.
<i>MaxError</i>	Maximum residual error for regression.
<i>MCCBinary</i>	Matthews correlation coefficient for binary classification.
<i>MCCMulticlass</i>	Matthews correlation coefficient for multiclass classification.
<i>MeanSquaredLogError</i>	Mean squared log error for regression.
<i>MedianAE</i>	Median absolute error for regression.
<i>MSE</i>	Mean squared error for regression.
<i>MulticlassClassificationObjective</i>	Base class for all multiclass classification objectives.
<i>ObjectiveBase</i>	Base class for all objectives.
<i>Precision</i>	Precision score for binary classification.
<i>PrecisionMacro</i>	Precision score for multiclass classification using macro averaging.
<i>PrecisionMicro</i>	Precision score for multiclass classification using micro averaging.

continues on next page

Table 246 – continued from previous page

<i>PrecisionWeighted</i>	Precision score for multiclass classification using weighted averaging.
<i>R2</i>	Coefficient of determination for regression.
<i>Recall</i>	Recall score for binary classification.
<i>RecallMacro</i>	Recall score for multiclass classification using macro averaging.
<i>RecallMicro</i>	Recall score for multiclass classification using micro averaging.
<i>RecallWeighted</i>	Recall score for multiclass classification using weighted averaging.
<i>RegressionObjective</i>	Base class for all regression objectives.
<i>RootMeanSquaredError</i>	Root mean squared error for regression.
<i>RootMeanSquaredLogError</i>	Root mean squared log error for regression.
<i>SensitivityLowAlert</i>	Base class for all binary classification objectives.

Functions

<i>get_all_objective_names</i>	Get a list of the names of all objectives.
<i>get_core_objective_names</i>	Get a list of all valid core objectives.
<i>get_core_objectives</i>	Returns all core objective instances associated with the given problem type.
<i>get_non_core_objectives</i>	Get non-core objective classes.
<i>get_objective</i>	Returns the Objective class corresponding to a given objective name.

Contents

class evalml.objectives.**AccuracyBinary**

Accuracy score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Accuracy Binary
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.

continues on next page

Table 248 – continued from previous page

<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: `y_predicted` (`pd.Series`): Predicted values of length `[n_samples]` `y_true` (`pd.Series`): Actual class labels of length `[n_samples]` `X` (`pd.DataFrame` or `np.ndarray`): Extra data of shape `[n_samples, n_features]` necessary to calculate score `sample_weight` (`pd.DataFrame` or `np.ndarray`): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

`optimize_threshold` (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **`ypred_proba`** (*pd.Series*) – The classifier’s predicted probabilities
- **`y_true`** (*pd.Series*) – The ground truth for the predictions.
- **`X`** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

`positive_only` (*cls*)

If True, this objective is only valid for positive data. Default False.

`score` (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **`y_predicted`** (*pd.Series*) – Predicted values of length `[n_samples]`
- **`y_true`** (*pd.Series*) – Actual class labels of length `[n_samples]`
- **`X`** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape `[n_samples, n_features]` necessary to calculate score
- **`sample_weight`** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

`validate_inputs` (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **`y_predicted`** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length `[n_samples]`
- **`y_true`** (*pd.Series*) – Actual class labels of length `[n_samples]`

Returns None

`class evalml.objectives.AccuracyMulticlass`

Accuracy score for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Accuracy Multiclass
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**AUC**

AUC score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC
per-fect_score	1.0
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	

continues on next page

Table 250 – continued from previous page

<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length `[n_samples]` *y_true* (*pd.Series*): Actual class labels of length `[n_samples]` *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape `[n_samples, n_features]` necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**AUCMacro**

AUC score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Macro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**AUCMicro**

AUC score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Micro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.

- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.**AUCWeighted**

AUC Score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	AUC Weighted
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**BalancedAccuracyBinary**

Balanced accuracy score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Balanced Accuracy Binary
per-fect_score	1.0
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	

continues on next page

Table 254 – continued from previous page

<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length `[n_samples]` *y_true* (*pd.Series*): Actual class labels of length `[n_samples]` *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape `[n_samples, n_features]` necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**BalancedAccuracyMulticlass**

Balanced accuracy score for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Balanced Accuracy Multiclass
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**BinaryClassificationObjective**

Base class for all binary classification objectives.

Attributes

problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
----------------------	--

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<i>greater_is_better</i>	Returns a boolean determining if a greater score indicates better model performance.
<i>is_bounded_like_percentage</i>	Returns whether this objective is bounded between 0 and 1, inclusive.
<i>is_defined_for_problem_type</i>	
<i>name</i>	Returns a name describing the objective.
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>perfect_score</i>	Returns the score obtained by evaluating this objective on a perfect model.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 256 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the <code>score()</code> method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property name (*cls*)

Returns a name describing the objective.

abstract classmethod objective_function (*cls, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

property perfect_score (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame or np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

property score_needs_proba (*cls*)

Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**CostBenefitMatrix**(*true_positive*, *true_negative*, *false_positive*, *false_negative*)

Score using a cost-benefit matrix. Scores quantify the benefits of a given value, so greater numeric scores represents a better score. Costs and scores can be negative, indicating that a value is not beneficial. For example, in the case of monetary profit, a negative cost and/or score represents loss of cash flow.

Parameters

- **true_positive** (*float*) – Cost associated with true positive predictions
- **true_negative** (*float*) – Cost associated with true negative predictions
- **false_positive** (*float*) – Cost associated with false positive predictions
- **false_negative** (*float*) – Cost associated with false negative predictions

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Cost Benefit Matrix
perfect_score	None
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculates cost-benefit of the using the predicted and true values.
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property can_optimize_threshold (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Calculates cost-benefit of the using the predicted and true values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted labels
- **y_true** (*pd.Series*) – True labels
- **X** (*pd.DataFrame*) – Ignored.
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns Cost-benefit matrix score

Return type float

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.

- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**ExpVariance**

Explained variance score for regression.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	ExpVariance
per-fect_score	1.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	

continues on next page

Table 258 – continued from previous page

<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**F1**

F1 score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In

practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold(cls)`

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score**validate_inputs** (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.**F1Macro**

F1 score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.F1Micro

F1 score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Micro
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of

shape [n_samples, n_features] necessary to calculate score sample_weight (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**F1Weighted**

F1 score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	F1 Weighted
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	

continues on next page

Table 262 – continued from previous page

<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**FraudCost** (*retry_percentage=0.5*, *interchange_fee=0.02*,
fraud_payout_percentage=1.0, *amount_col='amount'*)

Score the percentage of money lost of the total transaction amount process due to fraud.

Parameters

- **retry_percentage** (*float*) – What percentage of customers that will retry a transaction if it is declined. Between 0 and 1. Defaults to 0.5.
- **interchange_fee** (*float*) – How much of each successful transaction you pay. Between 0 and 1. Defaults to 0.02.
- **fraud_payout_percentage** (*float*) – Percentage of fraud you will not be able to collect. Between 0 and 1. Defaults to 1.0.
- **amount_col** (*str*) – Name of column in data that contains the amount. Defaults to “amount”.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	Fraud Cost
perfect_score	0.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculate amount lost to fraud per transaction given predictions, true values, and dataframe with transaction amount.
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.

continues on next page

Table 263 – continued from previous page

<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns `predictions`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X*, *sample_weight=None*)

Calculate amount lost to fraud per transaction given predictions, true values, and dataframe with transaction amount.

Parameters

- **y_predicted** (*pd.Series*) – Predicted fraud labels
- **y_true** (*pd.Series*) – True fraud labels
- **X** (*pd.DataFrame*) – Data with transaction amounts
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns `Amount lost to fraud per transaction`

Return type float

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

`evalml.objectives.get_all_objective_names()`

Get a list of the names of all objectives.

Returns Objective names

Return type list (str)

`evalml.objectives.get_core_objective_names()`

Get a list of all valid core objectives.

Returns Objective names.

Return type list[str]

`evalml.objectives.get_core_objectives(problem_type)`

Returns all core objective instances associated with the given problem type.

Core objectives are designed to work out-of-the-box for any dataset.

Parameters `problem_type` (*str/ProblemTypes*) – Type of problem

Returns List of ObjectiveBase instances

`evalml.objectives.get_non_core_objectives()`

Get non-core objective classes.

Non-core objectives are objectives that are domain-specific. Users typically need to configure these objectives before using them in AutoMLSearch.

Returns List of ObjectiveBase classes

`evalml.objectives.get_objective(objective, return_instance=False, **kwargs)`

Returns the Objective class corresponding to a given objective name.

Parameters

- **objective** (*str or ObjectiveBase*) – Name or instance of the objective class.
- **return_instance** (*bool*) – Whether to return an instance of the objective. This only applies if objective is of type *str*. Note that the instance will be initialized with default arguments.
- **kwargs** (*Any*) – Any keyword arguments to pass into the objective. Only used when `return_instance=True`.

Returns ObjectiveBase if the parameter objective is of type ObjectiveBase. If objective is instead a valid objective name, function will return the class corresponding to that name. If `return_instance` is True, an instance of that objective will be returned.

class `evalml.objectives.Gini`

Gini coefficient for binary classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Gini
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 264 – continued from previous page

<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**LeadScoring** (*true_positives=1*, *false_positives=-1*)

Lead scoring.

Parameters

- **true_positives** (*int*) – Reward for a true positive. Defaults to 1.
- **false_positives** (*int*) – Cost for a false positive. Should be negative. Defaults to -1.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	Lead Scoring
per-fect_score	None
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculate the profit per lead.
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series, np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float, optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Calculate the profit per lead.

Parameters

- **y_predicted** (*pd.Series*) – Predicted labels
- **y_true** (*pd.Series*) – True labels
- **X** (*pd.DataFrame*) – Ignored.
- **sample_weight** (*pd.DataFrame*) – Ignored.

Returns Profit per lead

Return type float

optimize_threshold (*self, ypred_proba, y_true, X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame, optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame or np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame or np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**LogLossBinary**

Log Loss for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	False
is_bounded_like_percentage	True
name	Log Loss Binary
perfect_score	0.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In

practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score**validate_inputs** (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None**class** evalml.objectives.**LogLossMulticlass**

Log Loss for multiclass classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	False
is_bounded_like_percentage	False
name	Log Loss Multiclass
perfect_score	0.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls, score, baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls, problem_type*)

objective_function (*self, y_true, y_predicted, X=None, sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self, y_true, y_predicted, X=None, sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self, y_true, y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.**MAE**
Mean absolute error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	MAE
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_positive	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of

shape [n_samples, n_features] necessary to calculate score sample_weight (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**MAPE**

Mean absolute percentage error for time series regression. Scaled by 100 to return a percentage.

Only valid for nonzero inputs. Otherwise, will throw a ValueError

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	True
name	Mean Absolute Percentage Error
per-fect_score	0.0
prob-lem_types	[ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	True

Methods

[*calculate_percent_difference*](#)

Calculate the percent difference between scores.

continues on next page

Table 269 – continued from previous page

<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod calculate_percent_difference (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score

- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**MaxError**

Maximum residual error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MaxError
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.MCCBinary

Matthews correlation coefficient for binary classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	MCC Binary
per-fect_score	1.0
prob-lem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property **can_optimize_threshold** (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**MCCMulticlass**

Matthews correlation coefficient for multiclass classification.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	MCC Multiclass
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- ***y_predicted*** (*pd.Series*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]
- ***X*** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- ***sample_weight*** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- ***y_predicted*** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.MeanSquaredLogError`

Mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a `ValueError`

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Mean Squared Log Error
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**MedianAE**

Median absolute error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MedianAE
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 274 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.MSE

Mean squared error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	MSE
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.**MulticlassClassificationObjective**

Base class for all multiclass classification objectives.

Attributes

prob- lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
----------------------------	--

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

continues on next page

Table 276 – continued from previous page

<code>greater_is_better</code>	Returns a boolean determining if a greater score indicates better model performance.
<code>is_bounded_like_percentage</code>	Returns whether this objective is bounded between 0 and 1, inclusive.
<code>is_defined_for_problem_type</code>	
<code>name</code>	Returns a name describing the objective.
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>perfect_score</code>	Returns the score obtained by evaluating this objective on a perfect model.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod objective_function (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property perfect_score (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property score_needs_proba (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.ObjectiveBase

Base class for all objectives.

Attributes

problem_types	None
---------------	------

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>expected_range</code>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].
<code>greater_is_better</code>	Returns a boolean determining if a greater score indicates better model performance.
<code>is_bounded_like_percentage</code>	Returns whether this objective is bounded between 0 and 1, inclusive.
<code>is_defined_for_problem_type</code>	
<code>name</code>	Returns a name describing the objective.
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>perfect_score</code>	Returns the score obtained by evaluating this objective on a perfect model.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod `objective_function` (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property `perfect_score` (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

property `positive_only` (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **`y_predicted`** (*pd.Series*) – Predicted values of length [*n_samples*]
- **`y_true`** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **`X`** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **`sample_weight`** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property `score_needs_proba` (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **`y_predicted`** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **`y_true`** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.Precision`

Precision score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Apply a learned threshold to predicted probabilities to get predicted classes.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [n_samples] *y_true* (*pd.Series*): Actual class labels of length [n_samples] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**PrecisionMacro**

Precision score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type(cls, problem_type)`

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- ***y_predicted*** (*pd.Series*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]
- ***X*** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- ***sample_weight*** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- ***y_predicted*** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- ***y_true*** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class `evalml.objectives.PrecisionMicro`

Precision score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Micro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.PrecisionWeighted

Precision score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Precision Weighted
per-fect_score	1.0
prob-lem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 281 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns `score`

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**R2**

Coefficient of determination for regression.

Attributes

ex-pected_range	None
greater_is_better	True
is_bounded_like_percentage	False
name	R2
per-fect_score	1
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.Recall

Recall score for binary classification.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>can_optimize_threshold</code>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<code>decision_function</code>	Apply a learned threshold to predicted probabilities to get predicted classes.
<code>is_defined_for_problem_type</code>	
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property `can_optimize_threshold` (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, *threshold=0.5*, *X=None*)

Apply a learned threshold to predicted probabilities to get predicted classes.

Parameters

- **ypred_proba** (*pd.Series*, *np.ndarray*) – The classifier’s predicted probabilities
- **threshold** (*float*, *optional*) – Threshold used to make a prediction. Defaults to 0.5.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns predictions

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.RecallMacro

Recall score for multiclass classification using macro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Macro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**RecallMicro**

Recall score for multiclass classification using micro averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Micro
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.

continues on next page

Table 285 – continued from previous page

<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns `score`

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.RecallWeighted

Recall score for multiclass classification using weighted averaging.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Recall Weighted
perfect_score	1.0
problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
score_needs_proba	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod **is_defined_for_problem_type** (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (pd.Series) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]
- **X** (pd.DataFrame or np.ndarray) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (pd.DataFrame or np.ndarray) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (pd.Series, or pd.DataFrame) – Predicted values of length [n_samples]
- **y_true** (pd.Series) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.RegressionObjective

Base class for all regression objectives.

Attributes

problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
----------------------	--

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>expected_range</i>	Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

continues on next page

Table 287 – continued from previous page

<code>greater_is_better</code>	Returns a boolean determining if a greater score indicates better model performance.
<code>is_bounded_like_percentage</code>	Returns whether this objective is bounded between 0 and 1, inclusive.
<code>is_defined_for_problem_type</code>	
<code>name</code>	Returns a name describing the objective.
<code>objective_function</code>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<code>perfect_score</code>	Returns the score obtained by evaluating this objective on a perfect model.
<code>positive_only</code>	If True, this objective is only valid for positive data. Default False.
<code>score</code>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<code>score_needs_proba</code>	Returns a boolean determining if the score() method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.
<code>validate_inputs</code>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type `float`

property `expected_range` (*cls*)

Returns the expected range of the objective, which is not necessarily the possible ranges. For example, our expected R2 range is from [-1, 1], although the actual range is (-inf, 1].

property `greater_is_better` (*cls*)

Returns a boolean determining if a greater score indicates better model performance.

property `is_bounded_like_percentage` (*cls*)

Returns whether this objective is bounded between 0 and 1, inclusive.

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

property `name` (*cls*)

Returns a name describing the objective.

abstract classmethod objective_function (*cls*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

property perfect_score (*cls*)

Returns the score obtained by evaluating this objective on a perfect model.

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

property score_needs_proba (*cls*)

Returns a boolean determining if the `score()` method needs probability estimates. This should be true for objectives which work with predicted probabilities, like log loss or AUC, and false for objectives which compare predicted class labels to the actual labels, like F1 or correlation.

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]

Returns None

class evalml.objectives.RootMeanSquaredError

Root mean squared error for regression.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Root Mean Squared Error
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (pd.Series): Predicted values of length [n_samples] *y_true* (pd.Series): Actual class labels of length [n_samples] *X* (pd.DataFrame or np.ndarray): Extra data of shape [n_samples, n_features] necessary to calculate score *sample_weight* (pd.DataFrame or np.ndarray): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**RootMeanSquaredLogError**

Root mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError.

Attributes

ex-pected_range	None
greater_is_better	False
is_bounded_like_percentage	False
name	Root Mean Squared Log Error
per-fect_score	0.0
prob-lem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	True

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

continues on next page

Table 289 – continued from previous page

<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod `calculate_percent_difference` (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

classmethod `is_defined_for_problem_type` (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Computes the relative value of the provided predictions compared to the actual labels, according a specified metric

Arguments: *y_predicted* (*pd.Series*): Predicted values of length [*n_samples*] *y_true* (*pd.Series*): Actual class labels of length [*n_samples*] *X* (*pd.DataFrame* or *np.ndarray*): Extra data of shape [*n_samples*, *n_features*] necessary to calculate score *sample_weight* (*pd.DataFrame* or *np.ndarray*): Sample weights used in computing objective value result

Returns Numerical value used to calculate score

positive_only (*self*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [*n_samples*]
- **y_true** (*pd.Series*) – Actual class labels of length [*n_samples*]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [*n_samples*, *n_features*] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

class evalml.objectives.**SensitivityLowAlert** (*alert_rate=0.01*)

Base class for all binary classification objectives.

Attributes

ex-pected_range	[0, 1]
greater_is_better	True
is_bounded_like_percentage	True
name	Sensitivity at Low Alert Rates
perfect_score	1.0
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_proba	False

Methods

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<i>can_optimize_threshold</i>	Returns a boolean determining if we can optimize the binary classification objective threshold.
<i>decision_function</i>	Determine if an observation is high risk given an alert rate
<i>is_defined_for_problem_type</i>	
<i>objective_function</i>	Calculate sensitivity across all predictions, using the top alert_rate percent of observations as the predicted positive class
<i>optimize_threshold</i>	Learn a binary classification threshold which optimizes the current objective.
<i>positive_only</i>	If True, this objective is only valid for positive data. Default False.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.
<i>validate_inputs</i>	Validates the input based on a few simple checks.

classmethod **calculate_percent_difference** (*cls*, *score*, *baseline_score*)

Calculate the percent difference between scores.

Parameters

- **score** (*float*) – A score. Output of the score method of this objective.
- **baseline_score** (*float*) – A score. Output of the score method of this objective. In practice, this is the score achieved on this objective with a baseline estimator.

Returns

The percent difference between the scores. Note that for objectives that can be interpreted as percentages, this will be the difference between the reference score and score. For all other objectives, the difference will be normalized by the reference score.

Return type float

property can_optimize_threshold (*cls*)

Returns a boolean determining if we can optimize the binary classification objective threshold. This will be false for any objective that works directly with predicted probabilities, like log loss and AUC. Otherwise, it will be true.

decision_function (*self*, *ypred_proba*, ***kwargs*)

Determine if an observation is high risk given an alert rate

Parameters *ypred_proba* (*pd.Series*) – Predicted probabilities

classmethod is_defined_for_problem_type (*cls*, *problem_type*)

objective_function (*self*, *y_true*, *y_predicted*, ***kwargs*)

Calculate sensitivity across all predictions, using the top alert_rate percent of observations as the predicted positive class

Parameters

- **y_true** (*pd.Series*) – True labels
- **y_predicted** (*pd.Series*) – Predicted labels based on alert_rate

Returns sensitivity using the observations with the top scores as the predicted positive class

Return type float

optimize_threshold (*self*, *ypred_proba*, *y_true*, *X=None*)

Learn a binary classification threshold which optimizes the current objective.

Parameters

- **ypred_proba** (*pd.Series*) – The classifier’s predicted probabilities
- **y_true** (*pd.Series*) – The ground truth for the predictions.
- **X** (*pd.DataFrame*, *optional*) – Any extra columns that are needed from training data.

Returns Optimal threshold for this objective

positive_only (*cls*)

If True, this objective is only valid for positive data. Default False.

score (*self*, *y_true*, *y_predicted*, *X=None*, *sample_weight=None*)

Returns a numerical score indicating performance based on the differences between the predicted and actual values.

Parameters

- **y_predicted** (*pd.Series*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]
- **X** (*pd.DataFrame* or *np.ndarray*) – Extra data of shape [n_samples, n_features] necessary to calculate score
- **sample_weight** (*pd.DataFrame* or *np.ndarray*) – Sample weights used in computing objective value result

Returns score

validate_inputs (*self*, *y_true*, *y_predicted*)

Validates the input based on a few simple checks.

Parameters

- **y_predicted** (*pd.Series*, or *pd.DataFrame*) – Predicted values of length [n_samples]
- **y_true** (*pd.Series*) – Actual class labels of length [n_samples]

Returns None

Pipelines

Subpackages

components

Subpackages

ensemble

Submodules

sklearn_stacked_ensemble_base

Module Contents

Classes Summary

SklearnStackedEnsembleBase

Stacked Ensemble Base Class.

Contents

class evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_base.**SklearnStackedEnsembleBase**

Stacked Ensemble Base Class.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.

- **final_estimator** (*Estimator or subclass*) – The estimator used to combine the base estimators.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, if the estimator is a classifier and y is either binary or multiclass, StratifiedKFold is used. In all other cases, KFold is used. Possible inputs for cv are:
 - None: 5-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
predict_uses_y	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path
<i>supported_problem_types</i>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

property supported_problem_types (*cls*)

Problem types this estimator supports

sklearn_stacked_ensemble_classifier

Module Contents

Classes Summary

<i>SklearnStackedEnsembleClassifier</i>

Scikit-learn Stacked Ensemble Classifier.

Contents

class evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.**SklearnStackedEnsembleClassifier**

Scikit-learn Stacked Ensemble Classifier.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be *None* or an empty list or else *EnsembleMissingPipelinesError* will be raised.
- **final_estimator** (*Estimator or subclass*) – The classifier used to combine the base estimators. If *None*, uses *LogisticRegressionClassifier*.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train *final_estimator*. For *int/None* inputs,

if the estimator is a classifier and `y` is either binary or multiclass, `StratifiedKFold` is used. Defaults to `None`. Possible inputs for `cv` are:

- `None`: 3-fold cross validation
- `int`: the number of folds in a (Stratified) `KFold`
- An scikit-learn cross-validation generator object
- An iterable yielding (train, test) splits
- **`n_jobs`** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. `None` and `1` are equivalent. If set to `-1`, all CPUs are used. For `n_jobs` below `-1`, $(n_cpus + 1 + n_jobs)$ are used. Defaults to `-1`. - Note: there could be some multi-process errors thrown for values of `n_jobs != 1`. If this is the case, please use `n_jobs = 1`.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to `0`.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Classifier
predict Uses y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for stacked ensemble classes.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Not implemented for <code>SklearnStackedEnsembleClassifier</code> and <code>SklearnStackedEnsembleRegressor</code>
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

sklearn_stacked_ensemble_regressor

Module Contents

Classes Summary

SklearnStackedEnsembleRegressor

Scikit-learn Stacked Ensemble Regressor.

Contents

class `evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_regressor.SklearnStackedEnsembleRegressor`

Scikit-learn Stacked Ensemble Regressor.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else `EnsembleMissingPipelinesError` will be raised.
- **final_estimator** (*Estimator or subclass*) – The regressor used to combine the base estimators. If None, uses `LinearRegressor`.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, `KFold` is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) `KFold`
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits

- **n_jobs** (*int* or *None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* $\neq 1$. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

Package Contents

Classes Summary

<i>SklearnStackedEnsembleBase</i>	Stacked Ensemble Base Class.
<i>SklearnStackedEnsembleClassifier</i>	Scikit-learn Stacked Ensemble Classifier.
<i>SklearnStackedEnsembleRegressor</i>	Scikit-learn Stacked Ensemble Regressor.

Contents

```
class evalml.pipelines.components.ensemble.SklearnStackedEnsembleBase (input_pipelines=None,
                                                                    fi-
                                                                    nal_estimator=None,
                                                                    cv=None,
                                                                    n_jobs=-
                                                                    1,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Stacked Ensemble Base Class.

Parameters

- **input_pipelines** (*list (PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The estimator used to combine the base estimators.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, if the estimator is a classifier and y is either binary or multiclass, StratifiedKFold is used. In all other cases, KFold is used. Possible inputs for cv are:
 - None: 5-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
pre-dict_uses_y	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path
<i>supported_problem_types</i>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

property supported_problem_types (*cls*)

Problem types this estimator supports

```

class evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier (input_pipelines=None,
final_estimator=None,
cv=None,
n_jobs=-1,
random_seed=0,
**kwargs)

```

Scikit-learn Stacked Ensemble Classifier.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The classifier used to combine the base estimators. If None, uses LogisticRegressionClassifier.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, if the estimator is a classifier and y is either binary or multiclass, StratifiedKFold is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for stacked ensemble classes.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor (input_pipelines=None,  
                                                                           fi-  
                                                                           nal_estimator=None,  
                                                                           cv=None,  
                                                                           n_jobs=-  
  
                                                                           1,  
                                                                           ran-  
                                                                           dom_seed=0,  
                                                                           **kwargs)
```

Scikit-learn Stacked Ensemble Regressor.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The regressor used to combine the base estimators. If None, uses LinearRegressor.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, KFold is used. Defaults to None. Possible inputs for cv are:

- None: 3-fold cross validation
- int: the number of folds in a (Stratified) KFold
- An scikit-learn cross-validation generator object
- An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* $\neq 1$. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

estimators

Subpackages

classifiers

Submodules

baseline_classifier

Module Contents

Classes Summary

BaselineClassifier

Classifier that predicts using the specified strategy.

Contents

class `evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClass`

Classifier that predicts using the specified strategy.

This is useful as a simple baseline classifier to compare with other classifiers.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mode”, “random” and “random_weighted”. Defaults to “mode”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS]

Methods

<i>classes_</i>	Returns class labels. Will return None before fitting.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

property *classes_*(*self*)

Returns class labels. Will return None before fitting.

Returns Class names

Return type list[str] or list(float)

clone(*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters(*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

catboost_classifier

Module Contents

Classes Summary

CatBoostClassifier

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.

Contents

class evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClasss

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.

- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Classifier
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

decision_tree_classifier

Module Contents

Classes Summary

<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
-------------------------------	---------------------------

Contents

class evalml.pipelines.components.estimators.classifiers.decision_tree_classifier.**DecisionTreeClassifier**

Decision Tree Classifier.

Parameters

- **criterion** (*{ "gini", "entropy" }*) – The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Defaults to “gini”.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and int(max_features * n_features) features are considered at each split.
 - If “auto”, then max_features=sqrt(n_features).
 - If “sqrt”, then max_features=sqrt(n_features).
 - If “log2”, then max_features=log2(n_features).
 - If None, then max_features = n_features.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.

- If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

Defaults to 2.

- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “criterion”: [“gini”, “entropy”], “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

`default_parameters` (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

elasticnet_classifier

Module Contents

Classes Summary

ElasticNetClassifier

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Contents

class evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.**ElasticNetC**

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Parameters

- **penalty** (*{ "l1", "l2", "elasticnet", "none" }*) – The norm used in penalization. Defaults to “elasticnet”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **multi_class** (*{ "auto", "ovr", "multinomial" }*) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when `solver='liblinear'`. “auto” selects “ovr” if the data is binary, or if `solver='liblinear'`, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (*{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }*) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good

choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.

- “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
- “liblinear” and “saga” also handle L1 penalty
- “saga” also supports “elasticnet” penalty
- “liblinear” does not support setting `penalty='none'`

Defaults to “saga”.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0.01, 10), “l1_ratio”: Real(0, 1)}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

et_classifier

Module Contents

Classes Summary

ExtraTreesClassifier

Extra Trees Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.et_classifier.**ExtraTreesClassifier**

Extra Trees Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int*, *float* or {"auto", "sqrt", "log2"}) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and *int(max_features * n_features)* features are considered at each split.
 - If “auto”, then *max_features*=*sqrt(n_features)*.
 - If “sqrt”, then *max_features*=*sqrt(n_features)*.

- If “log2”, then `max_features=log2(n_features)`.
- If None, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features. Defaults to “auto”.

- **`max_depth`** (*int*) – The maximum depth of the tree. Defaults to 6.
- **`min_samples_split`** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider `min_samples_split` as the minimum number.
 - If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.
- **`to 2.`** (*Defaults*) –
- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`n_jobs`** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10), }
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component

continues on next page

Table 310 – continued from previous page

<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

kneighbors_classifier

Module Contents

Classes Summary

KNeighborsClassifier

K-Nearest Neighbors Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.**KNeighborsC**

K-Nearest Neighbors Classifier.

Parameters

- **n_neighbors** (*int*) – Number of neighbors to use by default. Defaults to 5.

- **weights** ({ 'uniform', 'distance' } or callable) – Weight function used in prediction. Can be:
 - 'uniform' : uniform weights. All points in each neighborhood are weighted equally.
 - 'distance' : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
 - [callable] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.
 Defaults to “uniform”.
- **algorithm** ({ 'auto', 'ball_tree', 'kd_tree', 'brute' }) – Algorithm used to compute the nearest neighbors:
 - 'ball_tree' will use BallTree
 - 'kd_tree' will use KDTree
 - 'brute' will use a brute-force search.
 'auto' will attempt to decide the most appropriate algorithm based on the values passed to fit method. Defaults to “auto”. Note: fitting on sparse input will override the setting of this parameter, using brute force.
- **leaf_size** (int) – Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem. Defaults to 30.
- **p** (int) – Power parameter for the Minkowski metric. When $p = 1$, this is equivalent to using `manhattan_distance` (l1), and `euclidean_distance` (l2) for $p = 2$. For arbitrary p , `minkowski_distance (l_p)` is used. Defaults to 2.
- **random_seed** (int) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_neighbors": Integer(2, 12), "weights": ["uniform", "distance"], "algorithm": ["auto", "ball_tree", "kd_tree", "brute"], "leaf_size": Integer(10, 30), "p": Integer(1, 5), }
model_family	ModelFamily.K_NEIGHBORS
modifies_features	True
modifies_target	False
name	KNN Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters

continues on next page

Table 312 – continued from previous page

<code>feature_importance</code>	Returns array of 0's matching the input number of features as <code>feature_importance</code> is
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's matching the input number of features as `feature_importance` is not defined for KNN classifiers.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

lightgbm_classifier

Module Contents

Classes Summary

LightGBMClassifier

LightGBM Classifier.

Contents

`class evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier`

LightGBM Classifier.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select bagging_fraction * 100 % of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1), }
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Classifier
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

logistic_regression_classifier

Module Contents

Classes Summary

LogisticRegressionClassifier

Logistic Regression Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier. **Log**

Logistic Regression Classifier.

Parameters

- **penalty** ({*"l1"*, *"l2"*, *"elasticnet"*, *"none"*}) – The norm used in penalization. Defaults to *"l2"*.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **multi_class** ({*"auto"*, *"ovr"*, *"multinomial"*}) – If the option chosen is *"ovr"*, then a binary problem is fit for each label. For *"multinomial"* the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. *"multinomial"* is unavailable when solver=*"liblinear"*. *"auto"* selects *"ovr"* if the data is binary, or if solver=*"liblinear"*, and otherwise selects *"multinomial"*. Defaults to *"auto"*.
- **solver** ({*"newton-cg"*, *"lbfgs"*, *"liblinear"*, *"sag"*, *"saga"*}) – Algorithm to use in the optimization problem. For small datasets, *"liblinear"* is a good choice, whereas *"sag"* and *"saga"* are faster for large ones. For multiclass problems, only *"newton-cg"*, *"sag"*, *"saga"* and *"lbfgs"* handle multinomial loss; *"liblinear"* is limited to one-versus-rest schemes.
 - *"newton-cg"*, *"lbfgs"*, *"sag"* and *"saga"* handle L2 or no penalty
 - *"liblinear"* and *"saga"* also handle L1 penalty
 - *"saga"* also supports *"elasticnet"* penalty
 - *"liblinear"* does not support setting penalty=*'none'*
 Defaults to *"lbfgs"*.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "penalty": ["l2"], "C": Real(0.01, 10), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Logistic Regression Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

rf_classifier

Module Contents

Classes Summary

<i>RandomForestClassifier</i>

Random Forest Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.rf_classifier.**RandomForestClassifier**

Random Forest Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 10), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Classifier
predict Uses y	False
supported problem types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

svm_classifier

Module Contents

Classes Summary

SVMClassifier

Support Vector Machine Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.svm_classifier.**SVMClassifier** (*C=1.0*

ker-
nel='r
gamma
prob-
a-
bil-
ity=Tr
ran-
dom_s
**kwa

Support Vector Machine Classifier.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** ({*"poly"*, *"rbf"*, *"sigmoid"*}) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** ({*"scale"*, *"auto"*} or *float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses $1 / (n_features * X.var())$ as value of gamma - If “auto” (default), uses $1 / n_features$
- **probability** (*boolean*) – Whether to enable probability estimates. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.

continues on next page

Table 320 – continued from previous page

<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

xgboost_classifier**Module Contents****Classes Summary**

XGBoostClassifier

XGBoost Classifier.

Contents

class evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier

XGBoost Classifier.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ “eta”: Real(0.000001, 1), “max_depth”: Integer(1, 10), “min_child_weight”: Real(1, 10), “n_estimators”: Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Classifier
predict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

Package Contents

Classes Summary

<i>BaselineClassifier</i>	Classifier that predicts using the specified strategy.
<i>CatBoostClassifier</i>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.
<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
<i>ElasticNetClassifier</i>	Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.
<i>ExtraTreesClassifier</i>	Extra Trees Classifier.
<i>KNeighborsClassifier</i>	K-Nearest Neighbors Classifier.
<i>LightGBMClassifier</i>	LightGBM Classifier.
<i>LogisticRegressionClassifier</i>	Logistic Regression Classifier.
<i>RandomForestClassifier</i>	Random Forest Classifier.
<i>SVMClassifier</i>	Support Vector Machine Classifier.
<i>XGBoostClassifier</i>	XGBoost Classifier.

Contents

```
class evalml.pipelines.components.estimators.classifiers.BaselineClassifier (strategy='mode',  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Classifier that predicts using the specified strategy.

This is useful as a simple baseline classifier to compare with other classifiers.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mode”, “random” and “random_weighted”. Defaults to “mode”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS]

Methods

<i>classes_</i>	Returns class labels. Will return None before fitting.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

property *classes_*(*self*)

Returns class labels. Will return None before fitting.

Returns Class names

Return type list[str] or list(float)

clone(*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters(*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.CatBoostClassifier (n_estimators=10,
                                                                    eta=0.03,
                                                                    max_depth=6,
                                                                    boot-
                                                                    strap_type=None,
                                                                    silent=True,
                                                                    al-
                                                                    low_writing_files=Fa
                                                                    ran-
                                                                    dom_seed=0,
                                                                    n_jobs=-
                                                                    1,
                                                                    **kwargs)
```

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(4, 100), "eta": Real(0.000001, 1), "max_depth": Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier (criterion='gini',  
max_features=  
max_depth=6,  
min_samples_s  
min_weight_fra  
ran-  
dom_seed=0,  
**kwargs)
```

Decision Tree Classifier.

Parameters

- **criterion** (*{ "gini", "entropy" }*) – The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Defaults to “gini”.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “sqrt”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “log2”, then $\text{max_features} = \log_2(\text{n_features})$.
 - If None, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.
 - If float, then min_samples_split is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["gini", "entropy"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None


```

class evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier (penalty='elasticnet',
                                                                              C=1.0,
                                                                              l1_ratio=0.15,
                                                                              multi_class='auto',
                                                                              solver='saga',
                                                                              n_jobs=-1,
                                                                              random_seed=0,
                                                                              **kwargs)

```

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Parameters

- **penalty** (`{ "l1", "l2", "elasticnet", "none" }`) – The norm used in penalization. Defaults to “elasticnet”.
- **C** (`float`) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **l1_ratio** (`float`) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **multi_class** (`{ "auto", "ovr", "multinomial" }`) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when `solver="liblinear"`. “auto” selects “ovr” if the data is binary, or if `solver="liblinear"`, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (`{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }`) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.
 - “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
 - “liblinear” and “saga” also handle L1 penalty
 - “saga” also supports “elasticnet” penalty
 - “liblinear” does not support setting `penalty='none'`
 Defaults to “saga”.
- **n_jobs** (`int`) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (`int`) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0.01, 10), "l1_ratio": Real(0, 1)}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier (n_estimators=100,  
                                     max_features='auto',  
                                     max_depth=6,  
                                     min_samples_split=2,  
                                     min_weight_fraction_leaf=0.0,  
                                     n_jobs=-1,  
                                     random_state=None,  
                                     random_seed=0,  
                                     **kwargs)
```

Extra Trees Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “sqrt”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “log2”, then $\text{max_features} = \log_2(\text{n_features})$.
 - If *None*, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.
- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier (n_neighbors=5,  
                                                                    weights='uniform'  
                                                                    al-  
                                                                    go-  
                                                                    rithm='auto',  
                                                                    leaf_size=30,  
                                                                    p=2,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

K-Nearest Neighbors Classifier.

Parameters

- **n_neighbors** (*int*) – Number of neighbors to use by default. Defaults to 5.
- **weights** ({ *'uniform'*, *'distance'* } or *callable*) – Weight function used in prediction. Can be:
 - *'uniform'* : uniform weights. All points in each neighborhood are weighted equally.
 - *'distance'* : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
 - [*callable*] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Defaults to “uniform”.

- **algorithm** ({ *'auto'*, *'ball_tree'*, *'kd_tree'*, *'brute'* }) – Algorithm used to compute the nearest neighbors:
 - *'ball_tree'* will use BallTree
 - *'kd_tree'* will use KDTree
 - *'brute'* will use a brute-force search.

'auto' will attempt to decide the most appropriate algorithm based on the values passed to fit method. Defaults to “auto”. Note: fitting on sparse input will override the setting of this parameter, using brute force.
- **leaf_size** (*int*) – Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem. Defaults to 30.
- **p** (*int*) – Power parameter for the Minkowski metric. When $p = 1$, this is equivalent to using `manhattan_distance` (l1), and `euclidean_distance` (l2) for $p = 2$. For arbitrary p , `minkowski_distance` (l_p) is used. Defaults to 2.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_neighbors": Integer(2, 12), "weights": ["uniform", "distance"], "algorithm": ["auto", "ball_tree", "kd_tree", "brute"], "leaf_size": Integer(10, 30), "p": Integer(1, 5), }
model_family	ModelFamily.K_NEIGHBORS
modifies_features	True
modifies_target	False
name	KNN Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's matching the input number of features as feature_importance is
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's matching the input number of features as feature_importance is not defined for KNN classifiers.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.LightGBMClassifier (boosting_type='gbdt',  
learning_rate=0.1,  
n_estimators=100,  
max_depth=0,  
num_leaves=31,  
min_child_samples=20,  
bagging_fraction=0.9,  
bagging_freq=0,  
n_jobs=-1,  
random_seed=0,  
**kwargs)
```

LightGBM Classifier.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select bagging_fraction * 100 % of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1),}
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Classifier
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier (penalty,
C=1.0,
multi_
solver:
n_jobs

l,
ran-
dom_s
**kwa
```

Logistic Regression Classifier.

Parameters

- **penalty** ({*"l1"*, *"l2"*, *"elasticnet"*, *"none"*}) – The norm used in penalization. Defaults to *"l2"*.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **multi_class** ({*"auto"*, *"ovr"*, *"multinomial"*}) – If the option chosen is *"ovr"*, then a binary problem is fit for each label. For *"multinomial"* the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. *"multinomial"* is unavailable when *solver="liblinear"*. *"auto"* selects *"ovr"* if the data is binary, or if *solver="liblinear"*, and otherwise selects *"multinomial"*. Defaults to *"auto"*.
- **solver** ({*"newton-cg"*, *"lbfgs"*, *"liblinear"*, *"sag"*, *"saga"*}) – Algorithm to use in the optimization problem. For small datasets, *"liblinear"* is a good choice, whereas *"sag"* and *"saga"* are faster for large ones. For multiclass problems, only *"newton-cg"*, *"sag"*, *"saga"* and *"lbfgs"* handle multinomial loss; *"liblinear"* is limited to one-versus-rest schemes.
 - *"newton-cg"*, *"lbfgs"*, *"sag"* and *"saga"* handle L2 or no penalty
 - *"liblinear"* and *"saga"* also handle L1 penalty
 - *"saga"* also supports *"elasticnet"* penalty
 - *"liblinear"* does not support setting *penalty='none'*

Defaults to *"lbfgs"*.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "penalty": ["l2"], "C": Real(0.01, 10), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Logistic Regression Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```

class evalml.pipelines.components.estimators.classifiers.RandomForestClassifier(n_estimators=100,
                                                                              max_depth=6,
                                                                              n_jobs=-1,
                                                                              random_seed=0,
                                                                              **kwargs)

```

Random Forest Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 10), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (`pd.DataFrame`, or `np.ndarray`) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.classifiers.SVMClassifier (C=1.0,  
                                                                    ker-  
                                                                    nel='rbf',  
                                                                    gamma='auto',  
                                                                    prob-  
                                                                    a-  
                                                                    bil-  
                                                                    ity=True,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Support Vector Machine Classifier.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses 1 / (n_features * X.var()) as value of gamma - If “auto” (default), uses 1 / n_features
- **probability** (*boolean*) – Whether to enable probability estimates. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0, 10), "kernel": ["poly", "rbf", "sigmoid"], "gamma": ["scale", "auto"], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```

class evalml.pipelines.components.estimators.classifiers.XGBoostClassifier(eta=0.1,
                                                                           max_depth=6,
                                                                           min_child_weight=1,
                                                                           n_estimators=100,
                                                                           random_seed=0,
                                                                           eval_metric='logloss',
                                                                           n_jobs=12,
                                                                           **kwargs)

```

XGBoost Classifier.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 10), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Classifier
predict Uses y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path

continues on next page

Table 334 – continued from previous page

<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

regressors**Submodules****arima_regressor****Module Contents****Classes Summary**

ARIMARegressor

Autoregressive Integrated Moving Average Model.

Contents

class evalml.pipelines.components.estimators.regressors.arima_regressor.**ARIMAREgressor** (date_index, trend, start, d=0, start_q, max_p, max_d, max_q, seasonal, n_jobs, random_seed, **kwargs)

Autoregressive Integrated Moving Average Model. The three parameters (p, d, q) are the AR order, the degree of differencing, and the MA order. More information here: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

Currently ARIMAREgressor isn't supported via conda install. It's recommended that it be installed via PyPI.

Parameters

- **date_index** (*str*) – Specifies the name of the column in X that provides the datetime objects. Defaults to None.
- **trend** (*str*) – Controls the deterministic trend. Options are ['n', 'c', 't', 'ct'] where 'c' is a constant term, 't' indicates a linear trend, and 'ct' is both. Can also be an iterable when defining a polynomial, such as [1, 1, 0, 1].
- **start_p** (*int*) – Minimum Autoregressive order. Defaults to 2.
- **d** (*int*) – Minimum Differencing degree. Defaults to 0.
- **start_q** (*int*) – Minimum Moving Average order. Defaults to 2.
- **max_p** (*int*) – Maximum Autoregressive order. Defaults to 5.
- **max_d** (*int*) – Maximum Differencing degree. Defaults to 2.
- **max_q** (*int*) – Maximum Moving Average order. Defaults to 5.
- **seasonal** (*boolean*) – Whether to fit a seasonal model to ARIMA. Defaults to True.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "start_p": Integer(1, 3), "d": Integer(0, 2), "start_q": Integer(1, 3), "max_p": Integer(3, 10), "max_d": Integer(2, 5), "max_q": Integer(3, 10), "seasonal": [True, False], }
model_family	ModelFamily.ARIMA
modifies_features	True
modifies_target	False
name	ARIMA Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

baseline_regressor

Module Contents

Classes Summary

<i>BaselineRegressor</i>	Baseline regressor that uses a simple strategy to make predictions.
--------------------------	---

Contents

class evalml.pipelines.components.estimators.regressors.baseline_regressor.**BaselineRegressor**

Baseline regressor that uses a simple strategy to make predictions. This is useful as a simple baseline regressor to compare with other regressors.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mean”, “median”. Defaults to “mean”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data

continues on next page

Table 338 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

catboost_regressor**Module Contents****Classes Summary**

CatBoostRegressor

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.

Contents

class evalml.pipelines.components.estimators.regressors.catboost_regressor.CatBoostRegressor

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Regressor
predict Uses y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

decision_tree_regressor

Module Contents

Classes Summary

DecisionTreeRegressor

Decision Tree Regressor.

Contents

`class evalml.pipelines.components.estimators.regressors.decision_tree_regressor.DecisionTreeRegressor`

Decision Tree Regressor.

Parameters

- **criterion** (`{ "mse", "friedman_mse", "mae", "poisson" }`) – The function to measure the quality of a split. Supported criteria are:
 - “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node
 - “friedman_mse”, which uses mean squared error with Friedman’s improvement score for potential splits
 - “mae” for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,
 - “poisson” which uses reduction in Poisson deviance to find splits.
- **max_features** (`int, float or { "auto", "sqrt", "log2" }`) – The number of features to consider when looking for the best split:
 - If `int`, then consider `max_features` features at each split.
 - If `float`, then `max_features` is a fraction and `int(max_features * n_features)` features are considered at each split.
 - If “auto”, then `max_features=sqrt(n_features)`.
 - If “sqrt”, then `max_features=sqrt(n_features)`.
 - If “log2”, then `max_features=log2(n_features)`.
 - If `None`, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features.

- **max_depth** (`int`) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (`int or float`) – The minimum number of samples required to split an internal node:
 - If `int`, then consider `min_samples_split` as the minimum number.
 - If `float`, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (`float`) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (`int`) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["mse", "friedman_mse", "mae"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Regressor
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property `feature_importance` (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static `load` (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

elasticnet_regressor

Module Contents

Classes Summary

<i>ElasticNetRegressor</i>	Elastic Net Regressor.
----------------------------	------------------------

Contents

class evalml.pipelines.components.estimators.regressors.elasticnet_regressor.**ElasticNetReg**

Elastic Net Regressor.

Parameters

- **alpha** (*float*) – Constant that multiplies the penalty terms. Defaults to 0.0001.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **max_iter** (*int*) – The maximum number of iterations. Defaults to 1000.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. Defaults to False.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "alpha": Real(0, 1), "l1_ratio": Real(0, 1), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

et_regressor

Module Contents

Classes Summary

ExtraTreesRegressor

Extra Trees Regressor.

Contents

class evalml.pipelines.components.estimators.regressors.et_regressor.**ExtraTreesRegressor** (*n_*

Extra Trees Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “sqrt”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “log2”, then $\text{max_features} = \text{log2}(\text{n_features})$.
 - If *None*, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.
- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

lightgbm_regressor

Module Contents

Classes Summary

<i>LightGBMRegressor</i>	LightGBM Regressor.
--------------------------	---------------------

Contents

class evalml.pipelines.components.estimators.regressors.lightgbm_regressor.**LightGBMRegressor**

LightGBM Regressor.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select

bagging_fraction * 100 % of the data to use for the next k iterations. Defaults to 0.

- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “learning_rate”: Real(0.000001, 1), “boosting_type”: [“gbdt”, “dart”, “goss”, “rf”], “n_estimators”: Integer(10, 100), “max_depth”: Integer(0, 10), “num_leaves”: Integer(2, 100), “min_child_samples”: Integer(1, 100), “bagging_fraction”: Real(0.000001, 1), “bagging_freq”: Integer(0, 1),}
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Regressor
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

linear_regressor

Module Contents

Classes Summary

LinearRegressor

Linear Regressor.

Contents

class evalml.pipelines.components.estimators.regressors.linear_regressor.**LinearRegressor** (*fit*

Linear Regressor.

Parameters

- **fit_intercept** (*boolean*) – Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered). Defaults to True.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. This parameter is ignored when *fit_intercept* is set to False. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "fit_intercept": [True, False], "normalize": [True, False]}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Linear Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

prophet_regressor

Module Contents

Classes Summary

<i>ProphetRegressor</i>	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
-------------------------	--

Contents

class evalml.pipelines.components.estimators.regressors.prophet_regressor.**ProphetRegressor**

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

More information here: <https://facebook.github.io/prophet/>

Attributes

hyper-parameter_ranges	{ "changepoint_prior_scale": Real(0.001, 0.5), "seasonality_prior_scale": Real(0.01, 10), "holidays_prior_scale": Real(0.01, 10), "seasonality_mode": ["additive", "multiplicative"], }
model_family	ModelFamily.PROPHET
modifies_features	True
modifies_target	False
name	Prophet Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>build_prophet_df</i>	
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.
<i>fit</i>	Fits component to data
<i>get_params</i>	
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

static build_prophet_df (*X*, *y=None*, *date_column='ds'*)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

get_params (*self*)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X, y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

rf_regressor

Module Contents

Classes Summary

RandomForestRegressor

Random Forest Regressor.

Contents

class evalml.pipelines.components.estimators.regressors.rf_regressor.**RandomForestRegressor**

Random Forest Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_depth”: Integer(1, 32), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

svm_regressor

Module Contents

Classes Summary

SVMRegressor

Support Vector Machine Regressor.

Contents

```
class evalml.pipelines.components.estimators.regressors.svm_regressor.SVMRegressor (C=1.0,
                                                                    kernel='rbf',
                                                                    gamma='auto',
                                                                    random_seed=None,
                                                                    **kwargs)
```

Support Vector Machine Regressor.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses $1 / (n_features * X.var())$ as value of gamma - If “auto” (default), uses $1 / n_features$
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Regressor
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component

continues on next page

Table 356 – continued from previous page

<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

time_series_baseline_estimator

Module Contents

Classes Summary

<i>TimeSeriesBaselineEstimator</i>	Time series estimator that predicts using the naive forecasting approach.
------------------------------------	---

Contents

class evalml.pipelines.components.estimators.regressors.time_series_baseline_estimator.**Time**

Time series estimator that predicts using the naive forecasting approach.

This is useful as a simple baseline estimator for time series problems.

Parameters

- **gap** (*int*) – Gap between prediction date and target date and must be a positive integer. If gap is 0, target date will be shifted ahead by 1 time period. Defaults to 1.
- **forecast_horizon** (*int*) – Number of time steps the model is expected to predict.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Time Series Baseline Estimator
predict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property `feature_importance` (*self*)

Returns importance associated with each feature.

Since baseline estimators do not use input features to calculate predictions, returns an array of zeroes.

Returns an array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static `load` (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

xgboost_regressor

Module Contents

Classes Summary

XGBoostRegressor

XGBoost Regressor.

Contents

class evalml.pipelines.components.estimators.regressors.xgboost_regressor.XGBoostRegressor

XGBoost Regressor.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ “eta”: Real(0.000001, 1), “max_depth”: Integer(1, 20), “min_child_weight”: Real(1, 10), “n_estimators”: Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Regressor
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

Package Contents

Classes Summary

<i>ARIMAREgressor</i>	Autoregressive Integrated Moving Average Model.
<i>BaselineRegressor</i>	Baseline regressor that uses a simple strategy to make predictions.
<i>CatBoostRegressor</i>	CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.
<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>ElasticNetRegressor</i>	Elastic Net Regressor.
<i>ExtraTreesRegressor</i>	Extra Trees Regressor.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>LinearRegressor</i>	Linear Regressor.
<i>ProphetRegressor</i>	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
<i>RandomForestRegressor</i>	Random Forest Regressor.
<i>SVMRegressor</i>	Support Vector Machine Regressor.
<i>TimeSeriesBaselineEstimator</i>	Time series estimator that predicts using the naive forecasting approach.
<i>XGBoostRegressor</i>	XGBoost Regressor.

Contents

```
class evalml.pipelines.components.estimators.regressors.ARIMAREgressor (date_index=None,
                                                                    trend=None,
                                                                    start_p=2,
                                                                    d=0,
                                                                    start_q=2,
                                                                    max_p=5,
                                                                    max_d=2,
                                                                    max_q=5,
                                                                    seasonal=True,
                                                                    n_jobs=-1,
                                                                    random_seed=0,
                                                                    **kwargs)
```

Autoregressive Integrated Moving Average Model. The three parameters (p, d, q) are the AR order, the degree of differencing, and the MA order. More information here: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

Currently ARIMAREgressor isn't supported via conda install. It's recommended that it be installed via PyPI.

Parameters

- **date_index** (*str*) – Specifies the name of the column in X that provides the datetime objects. Defaults to None.
- **trend** (*str*) – Controls the deterministic trend. Options are ['n', 'c', 't', 'ct'] where 'c'

is a constant term, ‘t’ indicates a linear trend, and ‘ct’ is both. Can also be an iterable when defining a polynomial, such as [1, 1, 0, 1].

- **start_p**(*int*) – Minimum Autoregressive order. Defaults to 2.
- **d**(*int*) – Minimum Differencing degree. Defaults to 0.
- **start_q**(*int*) – Minimum Moving Average order. Defaults to 2.
- **max_p**(*int*) – Maximum Autoregressive order. Defaults to 5.
- **max_d**(*int*) – Maximum Differencing degree. Defaults to 2.
- **max_q**(*int*) – Maximum Moving Average order. Defaults to 5.
- **seasonal**(*boolean*) – Whether to fit a seasonal model to ARIMA. Defaults to True.
- **n_jobs**(*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to -1.
- **random_seed**(*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “start_p”: Integer(1, 3), “d”: Integer(0, 2), “start_q”: Integer(1, 3), “max_p”: Integer(3, 10), “max_d”: Integer(2, 5), “max_q”: Integer(3, 10), “seasonal”: [True, False], }
model_family	ModelFamily.ARIMA
modifies_features	True
modifies_target	False
name	ARIMA Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0’s with a length of 1 as feature_importance is not defined for ARIMA regressor.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone(*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.estimators.regressors.**BaselineRegressor** (*strategy='mean',*
random_seed=0,
***kwargs*)

Baseline regressor that uses a simple strategy to make predictions. This is useful as a simple baseline regressor to compare with other regressors.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mean”, “median”. Defaults to “mean”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

continues on next page

Table 363 – continued from previous page

<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.CatBoostRegressor (n_estimators=10,
                                                                           eta=0.03,
                                                                           max_depth=6,
                                                                           boot-
                                                                           strap_type=None,
                                                                           silent=False,
                                                                           al-
                                                                           low_writing_files=False,
                                                                           ran-
                                                                           dom_seed=0,
                                                                           n_jobs=-
                                                                           1,
                                                                           **kwargs)
```

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.

- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Regressor
predict Uses y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

class `evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor` (*criterion*='mse',
max_features='auto',
max_depth=6,
min_samples_split=2,
min_weight_fraction=0,
random_state=0,
***kwargs*)

Decision Tree Regressor.

Parameters

- **criterion** (`{ "mse", "friedman_mse", "mae", "poisson" }`) – The function to measure the quality of a split. Supported criteria are:
 - “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node
 - “friedman_mse”, which uses mean squared error with Friedman’s improvement score for potential splits
 - “mae” for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,
 - “poisson” which uses reduction in Poisson deviance to find splits.
- **max_features** (*int*, *float* or `{ "auto", "sqrt", "log2" }`) – The number of features to consider when looking for the best split:
 - If *int*, then consider `max_features` features at each split.
 - If *float*, then `max_features` is a fraction and `int(max_features * n_features)` features are considered at each split.
 - If “auto”, then `max_features=sqrt(n_features)`.
 - If “sqrt”, then `max_features=sqrt(n_features)`.
 - If “log2”, then `max_features=log2(n_features)`.
 - If *None*, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int* or *float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider `min_samples_split` as the minimum number.

- If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

Defaults to 2.

- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “criterion”: [“mse”, “friedman_mse”, “mae”], “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

`default_parameters` (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.ElasticNetRegressor (alpha=0.0001,  
                                                                    l1_ratio=0.15,  
                                                                    max_iter=1000,  
                                                                    nor-  
                                                                    mal-  
                                                                    ize=False,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Elastic Net Regressor.

Parameters

- **alpha** (*float*) – Constant that multiplies the penalty terms. Defaults to 0.0001.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **max_iter** (*int*) – The maximum number of iterations. Defaults to 1000.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. Defaults to False.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "alpha": Real(0, 1), "l1_ratio": Real(0, 1), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

clone

Constructs a new component with the same parameters and random state.

continues on next page

Table 366 – continued from previous page

<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor (n_estimators=100,
                                                                    max_features='auto',
                                                                    max_depth=6,
                                                                    min_samples_split=2,
                                                                    min_weight_fraction,
                                                                    n_jobs=-
                                                                    1,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Extra Trees Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int*, *float* *or* {"auto", "sqrt", "log2"}) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.

- If float, then `max_features` is a fraction and `int(max_features * n_features)` features are considered at each split.
- If “auto”, then `max_features=sqrt(n_features)`.
- If “sqrt”, then `max_features=sqrt(n_features)`.
- If “log2”, then `max_features=log2(n_features)`.
- If None, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features. Defaults to “auto”.

- **`max_depth`** (*int*) – The maximum depth of the tree. Defaults to 6.
- **`min_samples_split`** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider `min_samples_split` as the minimum number.
 - If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.
- **`to 2.`** (*Defaults*) –
- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`n_jobs`** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10), }
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data

continues on next page

Table 367 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.LightGBMRegressor (boosting_type='gbdt',  
                                                                    learn-  
                                                                    ing_rate=0.1,  
                                                                    n_estimators=20,  
                                                                    max_depth=0,  
                                                                    num_leaves=31,  
                                                                    min_child_samples=20,  
                                                                    bag-  
                                                                    ging_fraction=0.9,  
                                                                    bag-  
                                                                    ging_freq=0,  
                                                                    n_jobs=-  
  
                                                                    1,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

LightGBM Regressor.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest

- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, ≤ 0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select $\text{bagging_fraction} * 100 \%$ of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1), }
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Regressor
predict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.REGRESSION]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before

continues on next page

Table 368 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.estimators.regressors.LinearRegressor (fit_intercept=True,
                                                                    normalize=False,
                                                                    n_jobs=-1,
                                                                    random_seed=0,
                                                                    **kwargs)
```

Linear Regressor.

Parameters

- **fit_intercept** (*boolean*) – Whether to calculate the intercept for this model. If set to `False`, no intercept will be used in calculations (i.e. data is expected to be centered). Defaults to `True`.
- **normalize** (*boolean*) – If `True`, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. This parameter is ignored when `fit_intercept` is set to `False`. Defaults to `False`.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "fit_intercept": [True, False], "normalize": [True, False]}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Linear Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.ProphetRegressor (date_index=None,
                                                                           change-
                                                                           point_prior_scale=0.05,
                                                                           sea-
                                                                           son-
                                                                           al-
                                                                           ity_prior_scale=10,
                                                                           hol-
                                                                           i-
                                                                           days_prior_scale=10,
                                                                           sea-
                                                                           son-
                                                                           al-
                                                                           ity_mode='additive',
                                                                           ran-
                                                                           dom_seed=0,
                                                                           stan_backend='CMDSTAN',
                                                                           **kwargs)
```

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

More information here: <https://facebook.github.io/prophet/>

Attributes

hyper-parameter_ranges	{ "changepoint_prior_scale": Real(0.001, 0.5), "seasonality_prior_scale": Real(0.01, 10), "holidays_prior_scale": Real(0.01, 10), "seasonality_mode": ["additive", "multiplicative"], }
model_family	ModelFamily.PROPHET
modifies_features	True
modifies_target	False
name	Prophet Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>build_prophet_df</i>	
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.
<i>fit</i>	Fits component to data

continues on next page

Table 370 – continued from previous page

<i>get_params</i>	
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

static build_prophet_df (*X*, *y=None*, *date_column='ds'*)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

get_params (*self*)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.RandomForestRegressor (n_estimators=100,  
                                                                    max_depth=6,  
                                                                    n_jobs=-1,  
                                                                    1,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Random Forest Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 32), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None


```

class evalml.pipelines.components.estimators.regressors.SVMRegressor(C=1.0,
                                                                    ker-
                                                                    nel='rbf',
                                                                    gamma='auto',
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)

```

Support Vector Machine Regressor.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** ({*"poly"*, *"rbf"*, *"sigmoid"*}) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** ({*"scale"*, *"auto"*} or *float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses $1 / (n_features * X.var())$ as value of gamma - If “auto” (default), uses $1 / n_features$
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.

continues on next page

Table 372 – continued from previous page

<i>save</i>	Saves component at file path
clone (<i>self</i>)	
Constructs a new component with the same parameters and random state.	
Returns A new instance of this component with identical parameters and random state.	
default_parameters (<i>cls</i>)	
Returns the default parameters for this component.	
Our convention is that <code>Component.default_parameters == Component().parameters</code> .	
Returns default parameters for this component.	
Return type dict	
describe (<i>self</i> , <i>print_name=False</i> , <i>return_dict=False</i>)	
Describe a component and its parameters	
Parameters	
<ul style="list-style-type: none"> • print_name (<i>bool</i>, <i>optional</i>) – whether to print name of component • return_dict (<i>bool</i>, <i>optional</i>) – whether to return description as dictionary in the format {"name": name, "parameters": parameters} 	
Returns prints and returns dictionary	
Return type None or dict	
property feature_importance (<i>self</i>)	
Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros	
fit (<i>self</i> , <i>X</i> , <i>y=None</i>)	
Fits component to data	
Parameters	
<ul style="list-style-type: none"> • X (<i>list</i>, <i>pd.DataFrame</i> or <i>np.ndarray</i>) – The input training data of shape [n_samples, n_features] • y (<i>list</i>, <i>pd.Series</i>, <i>np.ndarray</i>, <i>optional</i>) – The target training data of length [n_samples] 	
Returns self	
static load (<i>file_path</i>)	
Loads component at file path	
Parameters file_path (<i>str</i>) – Location to load file	
Returns ComponentBase object	
needs_fitting (<i>self</i>)	
Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.	
property parameters (<i>self</i>)	
Returns the parameters which were used to initialize the component	
predict (<i>self</i> , <i>X</i>)	
Make predictions using selected features.	

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.estimators.regressors.**TimeSeriesBaselineEstimator** (*gap=1*,
forecast_horizon,
random_seed,
***kwargs*)

Time series estimator that predicts using the naive forecasting approach.

This is useful as a simple baseline estimator for time series problems.

Parameters

- **gap** (*int*) – Gap between prediction date and target date and must be a positive integer. If gap is 0, target date will be shifted ahead by 1 time period. Defaults to 1.
- **forecast_horizon** (*int*) – Number of time steps the model is expected to predict.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Time Series Baseline Estimator
predict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Since baseline estimators do not use input features to calculate predictions, returns an array of zeroes.

Returns an array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.regressors.XGBoostRegressor (eta=0.1,
                                                                    max_depth=6,
                                                                    min_child_weight=1,
                                                                    n_estimators=100,
                                                                    random_state=0,
                                                                    n_jobs=12,
                                                                    **kwargs)
```

XGBoost Regressor.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0

- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 20), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Regressor
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

Submodules

estimator

Module Contents

Classes Summary

<i>Estimator</i>	A component that fits and predicts given data.
------------------	--

Contents

class evalml.pipelines.components.estimators.estimator.**Estimator** (*parameters=None*,
component_obj=None,
random_seed=0,
***kwargs*)

A component that fits and predicts given data.

To implement a new Estimator, define your own class which is a subclass of Estimator, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Estimator component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
pre_dict_uses_y	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path
<i>supported_problem_types</i>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

property supported_problem_types (*cls*)

Problem types this estimator supports

Package Contents

Classes Summary

<i>ARIMAREgressor</i>	Autoregressive Integrated Moving Average Model.
<i>BaselineClassifier</i>	Classifier that predicts using the specified strategy.
<i>BaselineRegressor</i>	Baseline regressor that uses a simple strategy to make predictions.
<i>CatBoostClassifier</i>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.
<i>CatBoostRegressor</i>	CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.
<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>ElasticNetClassifier</i>	Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.
<i>ElasticNetRegressor</i>	Elastic Net Regressor.
<i>Estimator</i>	A component that fits and predicts given data.
<i>ExtraTreesClassifier</i>	Extra Trees Classifier.
<i>ExtraTreesRegressor</i>	Extra Trees Regressor.
<i>KNeighborsClassifier</i>	K-Nearest Neighbors Classifier.
<i>LightGBMClassifier</i>	LightGBM Classifier.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>LinearRegressor</i>	Linear Regressor.
<i>LogisticRegressionClassifier</i>	Logistic Regression Classifier.
<i>ProphetRegressor</i>	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
<i>RandomForestClassifier</i>	Random Forest Classifier.
<i>RandomForestRegressor</i>	Random Forest Regressor.
<i>SVMClassifier</i>	Support Vector Machine Classifier.
<i>SVMRegressor</i>	Support Vector Machine Regressor.
<i>TimeSeriesBaselineEstimator</i>	Time series estimator that predicts using the naive forecasting approach.
<i>XGBoostClassifier</i>	XGBoost Classifier.
<i>XGBoostRegressor</i>	XGBoost Regressor.

Contents

```
class evalml.pipelines.components.estimators.ARIMAREgressor (date_index=None,  
trend=None,  
start_p=2,  
d=0, start_q=2,  
max_p=5, max_d=2,  
max_q=5, seasonal=True,  
n_jobs=-1, random_seed=0,  
**kwargs)
```

Autoregressive Integrated Moving Average Model. The three parameters (p, d, q) are the AR order, the degree of differencing, and the MA order. More information here: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

Currently ARIMAREgressor isn't supported via conda install. It's recommended that it be installed via PyPI.

Parameters

- **date_index** (*str*) – Specifies the name of the column in X that provides the datetime objects. Defaults to None.
- **trend** (*str*) – Controls the deterministic trend. Options are ['n', 'c', 't', 'ct'] where 'c' is a constant term, 't' indicates a linear trend, and 'ct' is both. Can also be an iterable when defining a polynomial, such as [1, 1, 0, 1].
- **start_p** (*int*) – Minimum Autoregressive order. Defaults to 2.
- **d** (*int*) – Minimum Differencing degree. Defaults to 0.
- **start_q** (*int*) – Minimum Moving Average order. Defaults to 2.
- **max_p** (*int*) – Maximum Autoregressive order. Defaults to 5.
- **max_d** (*int*) – Maximum Differencing degree. Defaults to 2.
- **max_q** (*int*) – Maximum Moving Average order. Defaults to 5.
- **seasonal** (*boolean*) – Whether to fit a seasonal model to ARIMA. Defaults to True.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "start_p": Integer(1, 3), "d": Integer(0, 2), "start_q": Integer(1, 3), "max_p": Integer(3, 10), "max_d": Integer(2, 5), "max_q": Integer(3, 10), "seasonal": [True, False], }
model_family	ModelFamily.ARIMA
modifies_features	True
modifies_target	False
name	ARIMA Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.BaselineClassifier (strategy='mode',  
                                                             ran-  
                                                             dom_seed=0,  
                                                             **kwargs)
```

Classifier that predicts using the specified strategy.

This is useful as a simple baseline classifier to compare with other classifiers.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mode”, “random” and “random_weighted”. Defaults to “mode”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS]

Methods

<i>classes_</i>	Returns class labels. Will return None before fitting.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

property classes_ (*self*)

Returns class labels. Will return None before fitting.

Returns Class names

Return type list[str] or list(float)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.BaselineRegressor (strategy='mean',
                                                                ran-
                                                                dom_seed=0,
                                                                **kwargs)
```

Baseline regressor that uses a simple strategy to make predictions. This is useful as a simple baseline regressor to compare with other regressors.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mean”, “median”. Defaults to “mean”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

continues on next page

Table 380 – continued from previous page

<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.CatBoostClassifier (n_estimators=10,
                                                                eta=0.03,
                                                                max_depth=6,
                                                                boot-
                                                                strap_type=None,
                                                                silent=True,
                                                                al-
                                                                low_writing_files=False,
                                                                ran-
                                                                dom_seed=0,
                                                                n_jobs=- 1,
                                                                **kwargs)
```

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.

- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.estimators.CatBoostRegressor (n_estimators=10,  
                                                             eta=0.03,  
                                                             max_depth=6,  
                                                             boot-  
                                                             strap_type=None,  
                                                             silent=False, al-  
                                                             low_writing_files=False,  
                                                             ran-  
                                                             dom_seed=0,  
                                                             n_jobs=-1,  
                                                             **kwargs)
```

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None


```

class evalml.pipelines.components.estimators.DecisionTreeClassifier(criterion='gini',
                                                                    max_features='auto',
                                                                    max_depth=6,
                                                                    min_samples_split=2,
                                                                    min_weight_fraction_leaf=0.0,
                                                                    random_seed=0,
                                                                    **kwargs)

```

Decision Tree Classifier.

Parameters

- **criterion** (*{ "gini", "entropy" }*) – The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Defaults to “gini”.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and int(max_features * n_features) features are considered at each split.
 - If “auto”, then max_features=sqrt(n_features).
 - If “sqrt”, then max_features=sqrt(n_features).
 - If “log2”, then max_features=log2(n_features).
 - If None, then max_features = n_features.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.
 - If float, then min_samples_split is a fraction and ceil(min_samples_split * n_samples) are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["gini", "entropy"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.DecisionTreeRegressor (criterion='mse',  
                                                                    max_features='auto',  
                                                                    max_depth=6,  
                                                                    min_samples_split=2,  
                                                                    min_weight_fraction_leaf=0.0,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Decision Tree Regressor.

Parameters

- **criterion** (*{ "mse", "friedman_mse", "mae", "poisson" }*) – The function to measure the quality of a split. Supported criteria are:
 - “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node
 - “friedman_mse”, which uses mean squared error with Friedman’s improvement score for potential splits
 - “mae” for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,
 - “poisson” which uses reduction in Poisson deviance to find splits.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “sqrt”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “log2”, then $\text{max_features} = \text{log2}(\text{n_features})$.
 - If None, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.
 - If float, then min_samples_split is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["mse", "friedman_mse", "mae"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```

class evalml.pipelines.components.estimators.ElasticNetClassifier (penalty='elasticnet',
                                                                    C=1.0,
                                                                    l1_ratio=0.15,
                                                                    multi_class='auto',
                                                                    solver='saga',
                                                                    n_jobs=-
                                                                    1,          ran-
                                                                    dom_seed=0,
                                                                    **kwargs)

```

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Parameters

- **penalty** (`{ "l1", "l2", "elasticnet", "none" }`) – The norm used in penalization. Defaults to “elasticnet”.
- **C** (`float`) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **l1_ratio** (`float`) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **multi_class** (`{ "auto", "ovr", "multinomial" }`) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when `solver="liblinear"`. “auto” selects “ovr” if the data is binary, or if `solver="liblinear"`, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (`{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }`) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.
 - “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
 - “liblinear” and “saga” also handle L1 penalty
 - “saga” also supports “elasticnet” penalty
 - “liblinear” does not support setting `penalty='none'`
 Defaults to “saga”.
- **n_jobs** (`int`) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (`int`) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0.01, 10), "l1_ratio": Real(0, 1)}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.ElasticNetRegressor(alpha=0.0001,
                                                                l1_ratio=0.15,
                                                                max_iter=1000,
                                                                normalize=False,
                                                                random_seed=0,
                                                                **kwargs)
```

Elastic Net Regressor.

Parameters

- **alpha** (*float*) – Constant that multiplies the penalty terms. Defaults to 0.0001.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **max_iter** (*int*) – The maximum number of iterations. Defaults to 1000.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. Defaults to False.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "alpha": Real(0, 1), "l1_ratio": Real(0, 1), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component

continues on next page

Table 386 – continued from previous page

<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.estimators.**Estimator** (*parameters=None*, *component_obj=None*, *random_seed=0*, ***kwargs*)

A component that fits and predicts given data.

To implement a new Estimator, define your own class which is a subclass of Estimator, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Estimator component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
pre-dict_uses_y	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>name</code>	Returns string name of this component
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path
<code>supported_problem_types</code>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]

- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

property supported_problem_types (*cls*)

Problem types this estimator supports

```
class evalml.pipelines.components.estimators.ExtraTreesClassifier (n_estimators=100,  
                                                                max_features='auto',  
                                                                max_depth=6,  
                                                                min_samples_split=2,  
                                                                min_weight_fraction_leaf=0.0,  
                                                                n_jobs=-  
                                                                1,      ran-  
                                                                dom_seed=0,  
                                                                **kwargs)
```

Extra Trees Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider `max_features` features at each split.
 - If *float*, then `max_features` is a fraction and `int(max_features * n_features)` features are considered at each split.
 - If “auto”, then `max_features=sqrt(n_features)`.
 - If “sqrt”, then `max_features=sqrt(n_features)`.
 - If “log2”, then `max_features=log2(n_features)`.
 - If *None*, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider `min_samples_split` as the minimum number.
 - If *float*, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10), }
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.ExtraTreesRegressor (n_estimators=100,
                                                                max_features='auto',
                                                                max_depth=6,
                                                                min_samples_split=2,
                                                                min_weight_fraction_leaf=0.0,
                                                                n_jobs=-1,
                                                                random_seed=0,
                                                                **kwargs)
```

Extra Trees Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int*, *float* or {"auto", "sqrt", "log2"}) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and int(max_features * n_features) features are considered at each split.

- If “auto”, then `max_features=sqrt(n_features)`.
- If “sqrt”, then `max_features=sqrt(n_features)`.
- If “log2”, then `max_features=log2(n_features)`.
- If None, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features. Defaults to “auto”.

- **`max_depth`** (*int*) – The maximum depth of the tree. Defaults to 6.
- **`min_samples_split`** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider `min_samples_split` as the minimum number.
 - If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.
- **`to 2.`** (*Defaults*) –
- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`n_jobs`** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10), }
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path

continues on next page

Table 389 – continued from previous page

<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.KNeighborsClassifier (n_neighbors=5,  
                                                                weights='uniform',  
                                                                algo-  
                                                                rithm='auto',  
                                                                leaf_size=30,  
                                                                p=2, ran-  
                                                                dom_seed=0,  
                                                                **kwargs)
```

K-Nearest Neighbors Classifier.

Parameters

- **n_neighbors** (*int*) – Number of neighbors to use by default. Defaults to 5.
- **weights** ({*'uniform'*, *'distance'*} or *callable*) – Weight function used in prediction. Can be:
 - *'uniform'* : uniform weights. All points in each neighborhood are weighted equally.
 - *'distance'* : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
 - [*callable*] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Defaults to “uniform”.

- **algorithm** ({*'auto'*, *'ball_tree'*, *'kd_tree'*, *'brute'*}) – Algorithm used to compute the nearest neighbors:

- ‘ball_tree’ will use BallTree
- ‘kd_tree’ will use KDTree
- ‘brute’ will use a brute-force search.

‘auto’ will attempt to decide the most appropriate algorithm based on the values passed to fit method. Defaults to “auto”. Note: fitting on sparse input will override the setting of this parameter, using brute force.

- **leaf_size** (*int*) – Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem. Defaults to 30.
- **p** (*int*) – Power parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan_distance (l1), and euclidean_distance (l2) for $p = 2$. For arbitrary p , minkowski_distance (l_p) is used. Defaults to 2.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_neighbors”: Integer(2, 12), “weights”: [“uniform”, “distance”], “algorithm”: [“auto”, “ball_tree”, “kd_tree”, “brute”], “leaf_size”: Integer(10, 30), “p”: Integer(1, 5),}
model_family	ModelFamily.K_NEIGHBORS
modifies_features	True
modifies_target	False
name	KNN Classifier
predict Uses y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0’s matching the input number of features as feature_importance is
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's matching the input number of features as `feature_importance` is not defined for KNN classifiers.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.LightGBMClassifier (boosting_type='gbdt',
                                                                learn-
                                                                ing_rate=0.1,
                                                                n_estimators=100,
                                                                max_depth=0,
                                                                num_leaves=31,
                                                                min_child_samples=20,
                                                                bag-
                                                                ging_fraction=0.9,
                                                                bag-
                                                                ging_freq=0,
                                                                n_jobs=-
                                                                1,
                                                                ran-
                                                                dom_seed=0,
                                                                **kwargs)
```

LightGBM Classifier.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.

- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select `bagging_fraction * 100 %` of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1), }
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Classifier
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.estimators.LightGBMRegressor (boosting_type='gbdt',  
                                                             learn-  
                                                             ing_rate=0.1,  
                                                             n_estimators=20,  
                                                             max_depth=0,  
                                                             num_leaves=31,  
                                                             min_child_samples=20,  
                                                             bag-  
                                                             ging_fraction=0.9,  
                                                             bag-  
                                                             ging_freq=0,  
                                                             n_jobs=- 1, ran-  
                                                             dom_seed=0,  
                                                             **kwargs)
```

LightGBM Regressor.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select `bagging_fraction * 100 %` of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1),}
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Regressor
pre_dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.LinearRegressor (fit_intercept=True,  
normalize=False,  
n_jobs=-1, random_seed=0,  
**kwargs)
```

Linear Regressor.

Parameters

- **fit_intercept** (*boolean*) – Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered). Defaults to True.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. This parameter is ignored when fit_intercept is set to False. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "fit_intercept": [True, False], "normalize": [True, False]}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Linear Regressor
predict Uses y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.

continues on next page

Table 393 – continued from previous page

<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.LogisticRegressionClassifier (penalty='l2',
                                                                    C=1.0,
                                                                    multi_class='auto',
                                                                    solver='lbfgs',
                                                                    n_jobs=-1,
                                                                    random_state=0,
                                                                    **kwargs)
```

Logistic Regression Classifier.

Parameters

- **penalty** (*{ "l1", "l2", "elasticnet", "none" }*) – The norm used in penalization. Defaults to “l2”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **multi_class** (*{ "auto", "ovr", "multinomial" }*) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when solver=“liblinear”. “auto” selects “ovr” if the data is binary, or if solver=“liblinear”, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (*{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }*) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.
 - “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
 - “liblinear” and “saga” also handle L1 penalty
 - “saga” also supports “elasticnet” penalty

- “liblinear” does not support setting `penalty='none'`

Defaults to “lbfgs”.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “penalty”: [“l2”], “C”: Real(0.01, 10), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Logistic Regression Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.estimators.**ProphetRegressor** (*date_index*=None, *change-point_prior_scale*=0.05, *seasonality_prior_scale*=10, *holidays_prior_scale*=10, *seasonality_mode*='additive', *random_seed*=0, *stan_backend*='CMDSTANPY', ***kwargs*)

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

More information here: <https://facebook.github.io/prophet/>

Attributes

hyper-parameter_ranges	{ "changepoint_prior_scale": Real(0.001, 0.5), "seasonality_prior_scale": Real(0.01, 10), "holidays_prior_scale": Real(0.01, 10), "seasonality_mode": ["additive", "multiplicative"], }
model_family	ModelFamily.PROPHET
modifies_features	True
modifies_target	False
name	Prophet Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>build_prophet_df</i>	
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

continues on next page

Table 395 – continued from previous page

<i>fit</i>	Fits component to data
<i>get_params</i>	
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

static build_prophet_df (*X*, *y=None*, *date_column='ds'*)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

get_params (*self*)

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.RandomForestClassifier (n_estimators=100,  
                                                                max_depth=6,  
                                                                n_jobs=-  
                                                                1, ran-  
                                                                dom_seed=0,  
                                                                **kwargs)
```

Random Forest Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 10), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.RandomForestRegressor (n_estimators=100,  
                                                                max_depth=6,  
                                                                n_jobs=-  
                                                                1, random  
                                                                seed=0,  
                                                                **kwargs)
```

Random Forest Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 32), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Regressor
predict Uses y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.SVMClassifier (C=1.0, kernel='rbf',
                                                         gamma='auto',
                                                         probability=True,
                                                         random_seed=0,
                                                         **kwargs)
```

Support Vector Machine Classifier.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses 1 / (n_features * X.var()) as value of gamma - If “auto” (default), uses 1 / n_features
- **probability** (*boolean*) – Whether to enable probability estimates. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Feature importance only works with linear kernels.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.estimators.**SVMRegressor** (*C=1.0*, *kernel='rbf'*,
gamma='auto',
random_seed=0,
***kwargs*)

Support Vector Machine Regressor.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** ({*"poly"*, *"rbf"*, *"sigmoid"*}) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** ({*"scale"*, *"auto"*} or *float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses 1 / (n_features * X.var()) as value of gamma - If “auto” (default), uses 1 / n_features
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0, 10), "kernel": ["poly", "rbf", "sigmoid"], "gamma": ["scale", "auto"], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator(gap=1,
                                                                    fore-
                                                                    cast_horizon=1,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Time series estimator that predicts using the naive forecasting approach.

This is useful as a simple baseline estimator for time series problems.

Parameters

- **gap** (*int*) – Gap between prediction date and target date and must be a positive integer. If gap is 0, target date will be shifted ahead by 1 time period. Defaults to 1.
- **forecast_horizon** (*int*) – Number of time steps the model is expected to predict.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Time Series Baseline Estimator
predict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Since baseline estimators do not use input features to calculate predictions, returns an array of zeroes.

Returns an array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (`pd.DataFrame`, or `np.ndarray`) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.estimators.XGBoostClassifier(eta=0.1,  
                                                             max_depth=6,  
                                                             min_child_weight=1,  
                                                             n_estimators=100,  
                                                             ran-  
                                                             dom_seed=0,  
                                                             eval_metric='logloss',  
                                                             n_jobs=12,  
                                                             **kwargs)
```

XGBoost Classifier.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ “eta”: Real(0.000001, 1), “max_depth”: Integer(1, 10), “min_child_weight”: Real(1, 10), “n_estimators”: Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Classifier
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property `feature_importance` (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static `load` (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.estimators.XGBoostRegressor(eta=0.1,
                                                             max_depth=6,
                                                             min_child_weight=1,
                                                             n_estimators=100,
                                                             random_seed=0,
                                                             n_jobs=12,
                                                             **kwargs)
```

XGBoost Regressor.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 20), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Regressor
predict Uses y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 402 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transformers

Subpackages

dimensionality_reduction

Submodules

lda

Module Contents

Classes Summary

LinearDiscriminantAnalysis

Reduces the number of features by using Linear Discriminant Analysis.

Contents

class evalml.pipelines.components.transformers.dimensionality_reduction.lda.**LinearDiscrimi**

Reduces the number of features by using Linear Discriminant Analysis.

Parameters

- **n_components** (*int*) – The number of features to maintain after computation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Linear Discriminant Analysis Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)
Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

pca

Module Contents

Classes Summary

<i>PCA</i>	Reduces the number of features by using Principal Component Analysis (PCA).
------------	---

Contents

class `evalml.pipelines.components.transformers.dimensionality_reduction.pca.PCA` (*variance=0.95*, *n_components=*
ran-
dom_seed=0, ***kwargs*)

Reduces the number of features by using Principal Component Analysis (PCA).

Parameters

- **variance** (*float*) – The percentage of the original data variance that should be preserved when reducing the number of features. Defaults to 0.95.
- **n_components** (*int*) – The number of features to maintain after computing SVD. Defaults to None, but will override variance variable if set.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	Real(0.25, 1)}:type: {"variance"
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	PCA Transformer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

Package Contents

Classes Summary

<i>LinearDiscriminantAnalysis</i>	Reduces the number of features by using Linear Discriminant Analysis.
<i>PCA</i>	Reduces the number of features by using Principal Component Analysis (PCA).

Contents

class evalml.pipelines.components.transformers.dimensionality_reduction.**LinearDiscriminantAnalysis**

Reduces the number of features by using Linear Discriminant Analysis.

Parameters

- **n_components** (*int*) – The number of features to maintain after computation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Linear Discriminant Analysis Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.dimensionality_reduction.**PCA** (*variance=0.95*, *n_components=None*, *random_seed=0*, ***kwargs*)

Reduces the number of features by using Principal Component Analysis (PCA).

Parameters

- **variance** (*float*) – The percentage of the original data variance that should be preserved when reducing the number of features. Defaults to 0.95.
- **n_components** (*int*) – The number of features to maintain after computing SVD. Defaults to None, but will override variance variable if set.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	Real(0.25, 1)}:type: {"variance"}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	PCA Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

encoders

Submodules

onehot_encoder

Module Contents

Classes Summary

<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>OneHotEncoderMeta</i>	A version of the ComponentBaseMeta class which includes validation on an additional one-hot-encoder-specific method <i>categories</i> .

Contents

`class evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder` (*top_n=1, features_to_encode=None, categories=None, drop='if_binary', handle_unknown='ignore', handle_missing='error', random_seed=None, **kwargs*)

A transformer that encodes categorical features in a one-hot numeric array.

Parameters

- **top_n** (*int*) – Number of categories per column to encode. If *None*, all categories will be encoded. Otherwise, the *n* most frequent will be encoded and all others will be dropped. Defaults to 10.
- **features_to_encode** (*list[str]*) – List of columns to encode. All other columns will remain untouched. If *None*, all appropriate columns will be encoded. Defaults to *None*.
- **categories** (*list*) – A two dimensional list of categories, where *categories[i]* is a list of the categories for the column at index *i*. This can also be *None*, or “*auto*” if *top_n* is not *None*. Defaults to *None*.
- **drop** (*string, list*) – Method (“*first*” or “*if_binary*”) to use to drop one category per feature. Can also be a list specifying which categories to drop for each feature. Defaults to “*if_binary*”.
- **handle_unknown** (*string*) – Whether to ignore or error for unknown categories for a feature encountered during *fit* or *transform*. If either *top_n* or *categories* is used to limit the number of categories per column, this must be “*ignore*”. Defaults to “*ignore*”.
- **handle_missing** (*string*) – Options for how to handle missing (NaN) values encountered during *fit* or *transform*. If this is set to “*as_category*” and NaN values are within the *n* most frequent, “*nan*” values will be encoded as their own column. If this is set to “*error*”, any missing values encountered will raise an error. Defaults to “*error*”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	One Hot Encoder

Methods

<code>categories</code>	Returns a list of the unique categories to be encoded for the particular feature, in order.
<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the categorical features after fitting.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	One-hot encode the input data.

categories (*self*, *feature_name*)

Returns a list of the unique categories to be encoded for the particular feature, in order.

Parameters **feature_name** (*str*) – the name of any feature provided to one-hot encoder during fit

Returns the unique categories, in the same dtype as they were provided during fit

Return type `np.ndarray`

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type `dict`

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type `None` or `dict`

fit (*self*, X , *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self***fit_transform** (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***get_feature_names** (*self*)

Return feature names for the categorical features after fitting.

Feature names are formatted as {column name}_{category name}. In the event of a duplicate name, an integer will be added at the end of the feature name to distinguish it.

For example, consider a dataframe with a column called “A” and category “x_y” and another column called “A_x” with “y”. In this example, the feature names would be “A_x_y” and “A_x_y_1”.

Returns The feature names after encoding, provided in the same order as *input_features*.**Return type** *np.ndarray***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** *ComponentBase* object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None***transform** (*self*, *X*, *y=None*)

One-hot encode the input data.

Parameters

- **X** (*pd.DataFrame*) – Features to one-hot encode.
- **y** (*pd.Series*) – Ignored.

Returns Transformed data, where each categorical feature has been encoded into numerical columns using one-hot encoding.

Return type *pd.DataFrame*

class `evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoderMeta`
A version of the `ComponentBaseMeta` class which includes validation on an additional one-hot-encoder-specific method *categories*.

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	None
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<code>check_for_fit</code>	<i>check_for_fit</i> wraps a method that validates if <i>self._is_fitted</i> is <i>True</i> .
<code>register</code>	Register a virtual subclass of an ABC.
<code>set_fit</code>	

classmethod `check_for_fit` (*cls, method*)

check_for_fit wraps a method that validates if *self._is_fitted* is *True*. It raises an exception if *False* and calls and returns the wrapped method if *True*.

register (*cls, subclass*)

Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

classmethod `set_fit` (*cls, method*)

target_encoder

Module Contents

Classes Summary

<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.
----------------------	--

Contents

```
class evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder (cols=None, smoothing=1.0, handle_unknown='value', handle_missing='value', random_seed=None, **kwargs)
```

A transformer that encodes categorical features into target encodings.

Parameters

- **cols** (*list*) – Columns to encode. If None, all string columns will be encoded, otherwise only the columns provided will be encoded. Defaults to None
- **smoothing** (*float*) – The smoothing factor to apply. The larger this value is, the more influence the expected target value has on the resulting target encodings. Must be strictly larger than 0. Defaults to 1.0
- **handle_unknown** (*string*) – Determines how to handle unknown categories for a feature encountered. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **handle_missing** (*string*) – Determines how to handle missing values encountered during *fit* or *transform*. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Target Encoder

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Return feature names for the input features after fitting.

continues on next page

Table 414 – continued from previous page

<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Return feature names for the input features after fitting.

Returns The feature names after encoding

Return type np.array

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

Package Contents

Classes Summary

<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.

Contents

```
class evalml.pipelines.components.transformers.encoders.OneHotEncoder (top_n=10,
                                                                    fea-
                                                                    tures_to_encode=None,
                                                                    cate-
                                                                    gories=None,
                                                                    drop='if_binary',
                                                                    han-
                                                                    dle_unknown='ignore',
                                                                    han-
                                                                    dle_missing='error',
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

A transformer that encodes categorical features in a one-hot numeric array.

Parameters

- **top_n** (*int*) – Number of categories per column to encode. If *None*, all categories will be encoded. Otherwise, the *n* most frequent will be encoded and all others will be dropped. Defaults to 10.
- **features_to_encode** (*list[str]*) – List of columns to encode. All other columns will remain untouched. If *None*, all appropriate columns will be encoded. Defaults to *None*.
- **categories** (*list*) – A two dimensional list of categories, where *categories[i]* is a list of the categories for the column at index *i*. This can also be *None*, or “auto” if *top_n* is not *None*. Defaults to *None*.
- **drop** (*string, list*) – Method (“first” or “if_binary”) to use to drop one category per feature. Can also be a list specifying which categories to drop for each feature. Defaults to ‘if_binary’.
- **handle_unknown** (*string*) – Whether to ignore or error for unknown categories for a feature encountered during *fit* or *transform*. If either *top_n* or *categories* is used to limit the number of categories per column, this must be “ignore”. Defaults to “ignore”.
- **handle_missing** (*string*) – Options for how to handle missing (NaN) values encountered during *fit* or *transform*. If this is set to “as_category” and NaN values are within the *n* most frequent, “nan” values will be encoded as their own column. If this is set to “error”, any missing values encountered will raise an error. Defaults to “error”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	One Hot Encoder

Methods

<code>categories</code>	Returns a list of the unique categories to be encoded for the particular feature, in order.
<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the categorical features after fitting.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	One-hot encode the input data.

categories (*self*, *feature_name*)

Returns a list of the unique categories to be encoded for the particular feature, in order.

Parameters **feature_name** (*str*) – the name of any feature provided to one-hot encoder during fit

Returns the unique categories, in the same dtype as they were provided during fit

Return type np.ndarray

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Return feature names for the categorical features after fitting.

Feature names are formatted as {column name}_{category name}. In the event of a duplicate name, an integer will be added at the end of the feature name to distinguish it.

For example, consider a dataframe with a column called “A” and category “x_y” and another column called “A_x” with “y”. In this example, the feature names would be “A_x_y” and “A_x_y_1”.

Returns The feature names after encoding, provided in the same order as input_features.

Return type np.ndarray

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

One-hot encode the input data.

Parameters

- **X** (*pd.DataFrame*) – Features to one-hot encode.

- **y** (*pd.Series*) – Ignored.

Returns Transformed data, where each categorical feature has been encoded into numerical columns using one-hot encoding.

Return type `pd.DataFrame`

```
class evalml.pipelines.components.transformers.encoders.TargetEncoder (cols=None,  
                                                                    smooth-  
                                                                    ing=1.0,  
                                                                    han-  
                                                                    dle_unknown='value',  
                                                                    han-  
                                                                    dle_missing='value',  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

A transformer that encodes categorical features into target encodings.

Parameters

- **cols** (*list*) – Columns to encode. If `None`, all string columns will be encoded, otherwise only the columns provided will be encoded. Defaults to `None`
- **smoothing** (*float*) – The smoothing factor to apply. The larger this value is, the more influence the expected target value has on the resulting target encodings. Must be strictly larger than 0. Defaults to 1.0
- **handle_unknown** (*string*) – Determines how to handle unknown categories for a feature encountered. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **handle_missing** (*string*) – Determines how to handle missing values encountered during *fit* or *transform*. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	<code>{}</code>
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
name	Target Encoder

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data

continues on next page

Table 417 – continued from previous page

<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Return feature names for the input features after fitting.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

get_feature_names (*self*)

Return feature names for the input features after fitting.

Returns The feature names after encoding

Return type `np.array`

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns `ComponentBase` object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed *X*

Return type `pd.DataFrame`

feature_selection

Submodules

feature_selector

Module Contents

Classes Summary

<i>FeatureSelector</i>

Selects top features based on importance weights.

Contents

class evalml.pipelines.components.transformers.feature_selection.feature_selector.**FeatureS**

Selects top features based on importance weights.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=*None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

rf_classifier_feature_selector

Module Contents

Classes Summary

<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
------------------------------------	--

Contents

class `evalml.pipelines.components.transformers.feature_selection.rf_classifier_feature_sel`

Selects top features based on importance weights using a Random Forest classifier.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to -np.inf.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Classifier Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the <code>component_obj</code> does not have a <code>transform</code> method, will raise an <code>MethodPropertyNotFoundError</code> exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Transforms input data by selecting features. If the `component_obj` does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

rf_regressor_feature_selector

Module Contents

Classes Summary

RFRegressorSelectFromModel

Selects top features based on importance weights using a Random Forest regressor.

Contents

class evalml.pipelines.components.transformers.feature_selection.rf_regressor_feature_selection

Selects top features based on importance weights using a Random Forest regressor.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to -np.inf.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Regressor Select From Model

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_names</code>	Get names of selected features.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an <code>MethodPropertyNotFoundError</code> exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***get_names** (*self*)

Get names of selected features.

Returns List of the names of features selected**Return type** *list[str]***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** *ComponentBase* object**needs_fitting** (*self*)Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.**property parameters** (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None***transform** (*self, X, y=None*)Transforms input data by selecting features. If the *component_obj* does not have a *transform* method, will raise an *MethodPropertyNotFoundError* exception.**Parameters**

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data. Ignored.

Returns Transformed X**Return type** *pd.DataFrame*

Package Contents

Classes Summary

<i>FeatureSelector</i>	Selects top features based on importance weights.
<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.

Contents

class evalml.pipelines.components.transformers.feature_selection.**FeatureSelector** (*parameters=None, component_obj=None, random_seed=0, **kwargs*)

Selects top features based on importance weights.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 425 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an <code>MethodPropertyNotFoundError</code> exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type pd.DataFrame

class evalml.pipelines.components.transformers.feature_selection.RFClassifierSelectFromModel

Selects top features based on importance weights using a Random Forest classifier.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to `-np.inf`.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Classifier Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the <code>component_obj</code> does not have a <code>transform</code> method, will raise an <code>MethodPropertyNotFoundError</code> exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

class `evalml.pipelines.components.transformers.feature_selection.RFRegressorSelectFromModel`

Selects top features based on importance weights using a Random Forest regressor.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5. Defaults to `None`.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to `-np.inf`.

- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Regressor Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

imputers

Submodules

imputer

Module Contents

Classes Summary

<i>Imputer</i>	Imputes missing data according to a specified imputation strategy.
----------------	--

Contents

class `evalml.pipelines.components.transformers.imputers.imputer.Imputer` (*categorical_impute_strategy*, *categorical_fill_value=None*, *numeric_impute_strategy='most_frequent'*, *numeric_fill_value=None*, *random_seed=0*, ***kwargs*)

Imputes missing data according to a specified imputation strategy.

Parameters

- **categorical_impute_strategy** (*string*) – Impute strategy to use for string, object, boolean, categorical dtypes. Valid values include “most_frequent” and “constant”.
- **numeric_impute_strategy** (*string*) – Impute strategy to use for numeric columns. Valid values include “mean”, “median”, “most_frequent”, and “constant”.
- **categorical_fill_value** (*string*) – When `categorical_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of `None` will fill with the

string “missing_value”.

- **numeric_fill_value** (*int*, *float*) – When `numeric_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of `None` will fill with 0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “categorical_impute_strategy”: [“most_frequent”], “numeric_impute_strategy”: [“mean”, “median”, “most_frequent”], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by imputing missing values. ‘None’ values are converted to np.nan before imputation and are

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits imputer to data. 'None' values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by imputing missing values. 'None' values are converted to np.nan before imputation and are treated as the same.

Parameters

- **x** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X**Return type** *pd.DataFrame***per_column_imputer****Module Contents****Classes Summary**

<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
-------------------------	---

Contents

class evalml.pipelines.components.transformers.imputers.per_column_imputer.**PerColumnImputer**

Imputes missing data according to a specified imputation strategy per column.

Parameters

- **impute_strategies** (*dict*) – Column and {"impute_strategy": strategy, "fill_value":value} pairings. Valid values for impute strategy include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to None, which uses "most_frequent" for all columns. When impute_strategy == "constant", fill_value is used to replace missing data. When None, uses 0 when imputing numerical data and "missing_value" for strings or object data types.
- **default_impute_strategy** (*str*) – Impute strategy to fall back on when none is provided for a certain column. Valid values include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to "most_frequent".
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Per Column Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputers on input data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by imputing missing values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputers on input data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to fit.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms input data by imputing missing values.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to transform.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

simple_imputer

Module Contents

Classes Summary

<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
----------------------	--

Contents

class evalml.pipelines.components.transformers.imputers.simple_imputer.**SimpleImputer** (*impute_strategy*, *fill_value*, *random_seed*, *missing_value*, *random_state*, *dtype_backend*, *dtype_backend_kwargs*)

Imputes missing data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types.
- **fill_value** (*string*) – When impute_strategy == “constant”, fill_value is used to replace missing data. Defaults to 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Simple Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on X and transforms X

continues on next page

Table 433 – continued from previous page

<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – the target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=*None*)

Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type pd.DataFrame

target_imputer

Module Contents

Classes Summary

<i>TargetImputer</i>	Imputes missing target data according to a specified imputation strategy.
<i>TargetImputerMeta</i>	A version of the ComponentBaseMeta class which handles when input features is None

Contents

class evalml.pipelines.components.transformers.imputers.target_imputer.**TargetImputer** (*impute_*
fill_value
ran-
dom_see
***kwargs*

Imputes missing target data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types. Defaults to “most_frequent”.
- **fill_value** (*string*) – When impute_strategy == “constant”, fill_value is used to replace missing data. Defaults to None which uses 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Target Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to target data. ‘None’ values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on and transforms the input target data.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input target data by imputing missing values. ‘None’ and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits imputer to target data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]. Ignored.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples].

Returns self

fit_transform (*self*, *X*, *y*)

Fits on and transforms the input target data.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.
- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (pd.DataFrame, pd.Series)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*)
Transforms input target data by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.
- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (pd.DataFrame, pd.Series)

class evalml.pipelines.components.transformers.imputers.target_imputer.**TargetImputerMeta**
A version of the ComponentBaseMeta class which handles when input features is None

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	['predict', 'predict_proba', 'transform', 'inverse_transform']
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<i>check_for_fit</i>	<i>check_for_fit</i> wraps a method that validates if <i>self.is_fitted</i> is <i>True</i> .
<i>register</i>	Register a virtual subclass of an ABC.
<i>set_fit</i>	

classmethod check_for_fit (*cls*, *method*)
check_for_fit wraps a method that validates if *self.is_fitted* is *True*. It raises an exception if *False* and calls and returns the wrapped method if *True*.

register (*cls*, *subclass*)
Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

classmethod set_fit (*cls*, *method*)

Package Contents

Classes Summary

<i>Imputer</i>	Imputes missing data according to a specified imputation strategy.
<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>TargetImputer</i>	Imputes missing target data according to a specified imputation strategy.

Contents

class evalml.pipelines.components.transformers.imputers.**Imputer** (*categorical_impute_strategy='most_frequent', categorical_fill_value=None, numeric_impute_strategy='mean', numeric_fill_value=None, random_seed=0, **kwargs*)

Imputes missing data according to a specified imputation strategy.

Parameters

- **categorical_impute_strategy** (*string*) – Impute strategy to use for string, object, boolean, categorical dtypes. Valid values include “most_frequent” and “constant”.
- **numeric_impute_strategy** (*string*) – Impute strategy to use for numeric columns. Valid values include “mean”, “median”, “most_frequent”, and “constant”.
- **categorical_fill_value** (*string*) – When categorical_impute_strategy == “constant”, fill_value is used to replace missing data. The default value of None will fill with the string “missing_value”.
- **numeric_fill_value** (*int, float*) – When numeric_impute_strategy == “constant”, fill_value is used to replace missing data. The default value of None will fill with 0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "categorical_impute_strategy": ["most_frequent"], "numeric_impute_strategy": ["mean", "median", "most_frequent"], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. 'None' values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by imputing missing values. 'None' values are converted to np.nan before imputation and are

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame, np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by imputing missing values. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.components.transformers.imputers.PerColumnImputer (impute_strategies=None,
de-
fault_impute_strategy='m
ran-
dom_seed=0,
**kwargs)
```

Imputes missing data according to a specified imputation strategy per column.

Parameters

- **impute_strategies** (*dict*) – Column and {"impute_strategy": strategy, "fill_value":value} pairings. Valid values for impute strategy include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to None, which uses "most_frequent" for all columns. When impute_strategy == "constant", fill_value is used to replace missing data. When None, uses 0 when imputing numerical data and "missing_value" for strings or object data types.
- **default_impute_strategy** (*str*) – Impute strategy to fall back on when none is provided for a certain column. Valid values include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to "most_frequent".
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Per Column Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputers on input data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by imputing missing values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputers on input data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to fit.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by imputing missing values.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to transform.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.imputers.SimpleImputer (impute_strategy='most_frequent',  
                                                                    fill_value=None,  
                                                                    random_seed=0,  
                                                                    **kwargs)
```

Imputes missing data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types.
- **fill_value** (*string*) – When *impute_strategy* == “constant”, *fill_value* is used to replace missing data. Defaults to 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Simple Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are

continues on next page

Table 440 – continued from previous page

<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms input by imputing missing values. ‘None’ and <code>np.nan</code> values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, X , *y=None*)

Fits imputer to data. ‘None’ values are converted to `np.nan` before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [`n_samples`, `n_features`]
- **y** (*pd.Series*, *optional*) – the target training data of length [`n_samples`]

Returns *self*

fit_transform (*self*, X , *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type `pd.DataFrame`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns `ComponentBase` object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y=None*)

Transforms input by imputing missing values. ‘None’ and `np.nan` values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.components.transformers.imputers.TargetImputer (impute_strategy='most_frequent',
                                                                    fill_value=None,
                                                                    random_seed=0,
                                                                    **kwargs)
```

Imputes missing target data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types. Defaults to “most_frequent”.
- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to `None` which uses 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "impute_strategy": ["mean", "median", "most_frequent"] }
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Target Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to target data. 'None' values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on and transforms the input target data.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input target data by imputing missing values. 'None' and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits imputer to target data. ‘None’ values are converted to `np.nan` before imputation and are treated as the same.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – The input training data of shape `[n_samples, n_features]`. Ignored.
- **y** (`pd.Series`, optional) – The target training data of length `[n_samples]`.

Returns `self`

fit_transform (`self`, `X`, `y`)

Fits on and transforms the input target data.

Parameters

- **X** (`pd.DataFrame`) – Features. Ignored.
- **y** (`pd.Series`) – Target data to impute.

Returns The original X, transformed y

Return type (`pd.DataFrame`, `pd.Series`)

static load (`file_path`)

Loads component at file path

Parameters **file_path** (`str`) – Location to load file

Returns `ComponentBase` object

needs_fitting (`self`)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (`self`)

Returns the parameters which were used to initialize the component

save (`self`, `file_path`, `pickle_protocol=cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (`str`) – Location to save file
- **pickle_protocol** (`int`) – The pickle data stream format.

Returns `None`

transform (`self`, `X`, `y`)

Transforms input target data by imputing missing values. ‘None’ and `np.nan` values are treated as the same.

Parameters

- **X** (`pd.DataFrame`) – Features. Ignored.
- **y** (`pd.Series`) – Target data to impute.

Returns The original X, transformed y

Return type (`pd.DataFrame`, `pd.Series`)

preprocessing

Submodules

datetime_featurizer

Module Contents

Classes Summary

<code><i>DateTimeFeaturizer</i></code>	Transformer that can automatically extract features from datetime columns.
--	--

Contents

class evalml.pipelines.components.transformers.preprocessing.datetime_featurizer.**DateTimeFeaturizer**

Transformer that can automatically extract features from datetime columns.

Parameters

- **features_to_extract** (*list*) – List of features to extract. Valid options include “year”, “month”, “day_of_week”, “hour”. Defaults to None.
- **encode_as_categories** (*bool*) – Whether day-of-week and month features should be encoded as pandas “category” dtype. This allows OneHotEncoders to encode these features. Defaults to False.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DateTime Featurization Component

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Gets the categories of each datetime feature.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

get_feature_names (*self*)

Gets the categories of each datetime feature.

Returns Dictionary, where each key-value pair is a column name and a dictionary mapping the unique feature values to their integer encoding.

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

Parameters

- **x** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

delayed_feature_transformer

Module Contents

Classes Summary

<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
----------------------------------	--

Contents

class evalml.pipelines.components.transformers.preprocessing.delayed_feature_transformer.D

Transformer that delays input features and target variable for time series problems.

Parameters

- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **max_delay** (*int*) – Maximum number of time units to delay each feature. Defaults to 2.
- **forecast_horizon** (*int*) – The number of time periods the pipeline is expected to forecast.
- **delay_features** (*bool*) – Whether to delay the input features. Defaults to True.
- **delay_target** (*bool*) – Whether to delay the target. Defaults to True.
- **gap** (*int*) – The number of time units between when the features are collected and when the target is collected. For example, if you are predicting the next time step's target, gap=1. This is only needed because when gap=0, we need to be sure to start the lagging of the target variable at 1. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. This transformer performs the same regardless of the random seed provided.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Delayed Feature Transformer
needs_fitting	False
target_get_colname_prefix	target_delay_{ }

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DelayFeatureTransformer.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the delayed features for all features in X and y.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DelayFeatureTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, optional) – The target training data of length [n_samples]

Returns self**fit_transform** (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**property parameters** (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Computes the delayed features for all features in X and y.

For each feature in X, it will add a column to the output dataframe for each delay in the (inclusive) range [1, max_delay]. The values of each delayed feature are simply the original feature shifted forward in time by the delay amount. For example, a delay of 3 units means that the feature value at row n will be taken from the n-3rd row of that feature

If y is not None, it will also compute the delayed values for the target variable.

Parameters

- **X** (*pd.DataFrame* or *None*) – Data to transform. None is expected when only the target variable is being used.
- **y** (*pd.Series*, or *None*) – Target.

Returns Transformed X.**Return type** *pd.DataFrame*

drop_null_columns

Module Contents

Classes Summary

<i>DropNullColumns</i>	Transformer to drop features whose percentage of NaN values exceeds a specified threshold.
------------------------	--

Contents

class evalml.pipelines.components.transformers.preprocessing.drop_null_columns.**DropNullColumns**

Transformer to drop features whose percentage of NaN values exceeds a specified threshold.

Parameters

- **pct_null_threshold** (*float*) – The percentage of NaN values in an input feature to drop. Must be a value between [0, 1] inclusive. If equal to 0.0, will drop columns with any null values. If equal to 1.0, will drop columns with all null values. Defaults to 0.95.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Null Columns Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path

continues on next page

Table 447 – continued from previous page

<i>transform</i>	Transforms data X by dropping columns that exceed the threshold of null values.
------------------	---

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms data X by dropping columns that exceed the threshold of null values.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type pd.DataFrame

drop_rows_transformer

Module Contents

Classes Summary

<i>DropRowsTransformer</i>	Transformer to drop rows specified by row indices.
----------------------------	--

Contents

class evalml.pipelines.components.transformers.preprocessing.drop_rows_transformer.DropRowsTransformer

Transformer to drop rows specified by row indices.

Parameters

- **indices_to_drop** (*list*) – List of indices to drop in the input data. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Is not used by this component. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Drop Rows Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

featuretools

Module Contents

Classes Summary

<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
-----------------------	--

Contents

class evalml.pipelines.components.transformers.preprocessing.featuretools.**DFSTransformer** (in ra
da
*)

Featuretools DFS component that generates features for the input features.

Parameters

- **index** (*string*) – The name of the column that contains the indices. If no column with this name exists, then featuretools.EntitySet() creates a column with this name to serve as the index column. Defaults to 'index'.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DFSTransformer Transformer component.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path

continues on next page

Table 451 – continued from previous page

<i>transform</i>	Computes the feature matrix for the input X using featuretools' dfs algorithm.
<hr/>	
clone (<i>self</i>)	
	Constructs a new component with the same parameters and random state.
	Returns A new instance of this component with identical parameters and random state.
default_parameters (<i>cls</i>)	
	Returns the default parameters for this component.
	Our convention is that <code>Component.default_parameters == Component().parameters</code> .
	Returns default parameters for this component.
	Return type dict
describe (<i>self</i> , <i>print_name=False</i> , <i>return_dict=False</i>)	
	Describe a component and its parameters
	Parameters
	<ul style="list-style-type: none"> • print_name (<i>bool</i>, <i>optional</i>) – whether to print name of component • return_dict (<i>bool</i>, <i>optional</i>) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}
	Returns prints and returns dictionary
	Return type None or dict
fit (<i>self</i> , <i>X</i> , <i>y=None</i>)	
	Fits the DFSTransformer Transformer component.
	Parameters
	<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i>, <i>np.array</i>) – The input data to transform, of shape [n_samples, n_features] • y (<i>pd.Series</i>, <i>np.ndarray</i>, <i>optional</i>) – The target training data of length [n_samples]
	Returns self
fit_transform (<i>self</i> , <i>X</i> , <i>y=None</i>)	
	Fits on X and transforms X
	Parameters
	<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i>) – Data to fit and transform • y (<i>pd.Series</i>) – Target data
	Returns Transformed X
	Return type pd.DataFrame
static load (<i>file_path</i>)	
	Loads component at file path
	Parameters file_path (<i>str</i>) – Location to load file
	Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Computes the feature matrix for the input X using featuretools' dfs algorithm.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data to transform. Has shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Feature matrix

Return type pd.DataFrame

log_transformer**Module Contents****Classes Summary**

LogTransformer

Applies a log transformation to the target data.

Contents

class evalml.pipelines.components.transformers.preprocessing.log_transformer.**LogTransformer**

Applies a log transformation to the target data.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Log Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the LogTransformer.
<i>fit_transform</i>	Log transforms the target variable.
<i>inverse_transform</i>	Inverts the transformation done by the transform method.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Log transforms the target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the LogTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Ignored.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to log transform.

Returns

The input features are returned without modification. The target variable *y* is log transformed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Inverts the transformation done by the transform method.

Arguments: *y* (*pd.Series*): Target transformed by this component.

Returns Target without the transformation.

Return type *pd.Series*∅

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target data to log transform.

Returns

The input features are returned without modification. The target variable y is log transformed.

Return type tuple of pd.DataFrame, pd.Series

lsa

Module Contents

Classes Summary

<i>LSA</i>	Transformer to calculate the Latent Semantic Analysis Values of text input.
------------	---

Contents

class evalml.pipelines.components.transformers.preprocessing.lsa.**LSA**(*random_seed=0*,
***kwargs*)

Transformer to calculate the Latent Semantic Analysis Values of text input.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	LSA Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path

continues on next page

Table 455 – continued from previous page

<i>transform</i>	Transforms data X by applying the LSA pipeline.
clone (<i>self</i>)	
Constructs a new component with the same parameters and random state.	
Returns A new instance of this component with identical parameters and random state.	
default_parameters (<i>cls</i>)	
Returns the default parameters for this component.	
Our convention is that <code>Component.default_parameters == Component().parameters</code> .	
Returns default parameters for this component.	
Return type dict	
describe (<i>self</i> , <i>print_name=False</i> , <i>return_dict=False</i>)	
Describe a component and its parameters	
Parameters	
<ul style="list-style-type: none"> • print_name (<i>bool</i>, <i>optional</i>) – whether to print name of component • return_dict (<i>bool</i>, <i>optional</i>) – whether to return description as dictionary in the format {"name": name, "parameters": parameters} 	
Returns prints and returns dictionary	
Return type None or dict	
fit (<i>self</i> , <i>X</i> , <i>y=None</i>)	
Fits component to data	
Parameters	
<ul style="list-style-type: none"> • X (<i>list</i>, <i>pd.DataFrame</i> or <i>np.ndarray</i>) – The input training data of shape [n_samples, n_features] • y (<i>list</i>, <i>pd.Series</i>, <i>np.ndarray</i>, <i>optional</i>) – The target training data of length [n_samples] 	
Returns self	
fit_transform (<i>self</i> , <i>X</i> , <i>y=None</i>)	
Fits on X and transforms X	
Parameters	
<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i>) – Data to fit and transform • y (<i>pd.Series</i>) – Target data 	
Returns Transformed X	
Return type pd.DataFrame	
static load (<i>file_path</i>)	
Loads component at file path	
Parameters file_path (<i>str</i>) – Location to load file	
Returns ComponentBase object	

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by applying the LSA pipeline.

Parameters

- **X** (*pd.DataFrame*) – The data to transform.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns

Transformed X. The original column is removed and replaced with two columns of the format *LSA(original_column_name)[feature_number]*, where *feature_number* is 0 or 1.

Return type pd.DataFrame

polynomial_detrender

Module Contents

Classes Summary

PolynomialDetrender

Removes trends from time series by fitting a polynomial to the data.

Contents

class evalml.pipelines.components.transformers.preprocessing.polynomial_detrender.**Polynomi**

Removes trends from time series by fitting a polynomial to the data.

Parameters

- **degree** (*int*) – Degree for the polynomial. If 1, linear model is fit to the data. If 2, quadratic model is fit, etc. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “degree”: Integer(1, 3)}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Polynomial Detrender

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the PolynomialDetrender.
<i>fit_transform</i>	Removes fitted trend from target variable.
<i>inverse_transform</i>	Adds back fitted trend to target variable.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Removes fitted trend from target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the PolynomialDetrender.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The first element are the input features returned without modification. The second element is the target variable *y* with the fitted trend removed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Adds back fitted trend to target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable.

Returns

The first element are the input features returned without modification. The second element is the target variable *y* with the trend added back.

Return type tuple of *pd.DataFrame*, *pd.Series*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **x** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The input features are returned without modification. The target variable *y* is detrended

Return type tuple of *pd.DataFrame*, *pd.Series*

text_featurizer

Module Contents

Classes Summary

TextFeaturizer

Transformer that can automatically featurize text columns using featuretools' *nlp_primitives*.

Contents

class evalml.pipelines.components.transformers.preprocessing.text_featurizer.**TextFeaturizer**

Transformer that can automatically featurize text columns using featuretools' *nlp_primitives*.

Since models cannot handle non-numeric data, any text must be broken down into features that provide useful information about that text. This component splits each text column into several informative features: Diversity Score, Mean Characters per Word, Polarity Score, and LSA (Latent Semantic Analysis). Calling transform on this component will replace any text columns in the given dataset with these numeric columns.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Text Featurization Component

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by creating new features using existing text columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self, X, y=None*)

Transforms data X by creating new features using existing text columns

Parameters

- **X** (*pd.DataFrame*) – The data to transform.
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

text_transformer

Module Contents

Classes Summary

TextTransformer

Base class for all transformers working with text features.

Contents

class evalml.pipelines.components.transformers.preprocessing.text_transformer.**TextTransform**

Base class for all transformers working with text features.

Parameters

- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

transform_primitive_components

Module Contents

Classes Summary

<i>EmailFeaturizer</i>	Transformer that can automatically extract features from emails.
<i>URLFeaturizer</i>	Transformer that can automatically extract features from URL.

Contents

class evalml.pipelines.components.transformers.preprocessing.transform_primitive_component.

Transformer that can automatically extract features from emails.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Email Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path

continues on next page

Table 463 – continued from previous page

<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

class evalml.pipelines.components.transformers.preprocessing.transform_primitive_components.

Transformer that can automatically extract features from URL.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	URL Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X

continues on next page

Table 464 – continued from previous page

<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=*None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

Package Contents

Classes Summary

<i>DateTimeFeaturizer</i>	Transformer that can automatically extract features from datetime columns.
<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
<i>DropNullColumns</i>	Transformer to drop features whose percentage of NaN values exceeds a specified threshold.
<i>DropRowsTransformer</i>	Transformer to drop rows specified by row indices.
<i>EmailFeaturizer</i>	Transformer that can automatically extract features from emails.
<i>LogTransformer</i>	Applies a log transformation to the target data.
<i>LSA</i>	Transformer to calculate the Latent Semantic Analysis Values of text input.
<i>PolynomialDetrender</i>	Removes trends from time series by fitting a polynomial to the data.

continues on next page

Table 465 – continued from previous page

<i>TextFeaturizer</i>	Transformer that can automatically featurize text columns using featuretools’ nlp_primitives.
<i>TextTransformer</i>	Base class for all transformers working with text features.
<i>URLFeaturizer</i>	Transformer that can automatically extract features from URL.

Contents

class evalml.pipelines.components.transformers.preprocessing.**DateTimeFeaturizer** (*features_to_extract*, *encode_as_categories*, *date_index=None*, *random_seed=0*, ***kwargs*)

Transformer that can automatically extract features from datetime columns.

Parameters

- **features_to_extract** (*list*) – List of features to extract. Valid options include “year”, “month”, “day_of_week”, “hour”. Defaults to None.
- **encode_as_categories** (*bool*) – Whether day-of-week and month features should be encoded as pandas “category” dtype. This allows OneHotEncoders to encode these features. Defaults to False.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DateTime Featurization Component

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Gets the categories of each datetime feature.

continues on next page

Table 466 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Gets the categories of each datetime feature.

Returns Dictionary, where each key-value pair is a column name and a dictionary mapping the unique feature values to their integer encoding.

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.preprocessing.**DelayedFeatureTransformer** (*date_*
max_
gap=
fore-
cast_
de-
lay_f
de-
lay_t
ran-
dom_
***kw*

Transformer that delays input features and target variable for time series problems.

Parameters

- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.

- **max_delay** (*int*) – Maximum number of time units to delay each feature. Defaults to 2.
- **forecast_horizon** (*int*) – The number of time periods the pipeline is expected to forecast.
- **delay_features** (*bool*) – Whether to delay the input features. Defaults to True.
- **delay_target** (*bool*) – Whether to delay the target. Defaults to True.
- **gap** (*int*) – The number of time units between when the features are collected and when the target is collected. For example, if you are predicting the next time step’s target, gap=1. This is only needed because when gap=0, we need to be sure to start the lagging of the target variable at 1. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. This transformer performs the same regardless of the random seed provided.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Delayed Feature Transformer
needs_fitting	False
target_colname_prefix	target_delay_{ }

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DelayFeatureTransformer.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the delayed features for all features in X and y.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DelayFeatureTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Computes the delayed features for all features in X and y.

For each feature in X, it will add a column to the output dataframe for each delay in the (inclusive) range [1, max_delay]. The values of each delayed feature are simply the original feature shifted forward in time by the delay amount. For example, a delay of 3 units means that the feature value at row n will be taken from the n-3rd row of that feature

If y is not None, it will also compute the delayed values for the target variable.

Parameters

- **X** (*pd.DataFrame* or *None*) – Data to transform. None is expected when only the target variable is being used.
- **y** (*pd.Series*, or *None*) – Target.

Returns Transformed X.

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.preprocessing.DFSTransformer (index='index',  
ran-  
dom_seed=0,  
**kwargs)
```

Featuretools DFS component that generates features for the input features.

Parameters

- **index** (*string*) – The name of the column that contains the indices. If no column with this name exists, then featuretools.EntitySet() creates a column with this name to serve as the index column. Defaults to 'index'.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DFSTransformer Transformer component.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component

continues on next page

Table 468 – continued from previous page

<i>save</i>	Saves component at file path
<i>transform</i>	Computes the feature matrix for the input X using featuretools’ dfs algorithm.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DFSTransformer Transformer component.

Parameters

- **X** (*pd.DataFrame*, *np.array*) – The input data to transform, of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=*None*)

Computes the feature matrix for the input X using featuretools’ dfs algorithm.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data to transform. Has shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Feature matrix

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.preprocessing.**DropNullColumns** (*pct_null_threshold*=*0.95*, *random_seed*=*0*, ***kwargs*)

Transformer to drop features whose percentage of NaN values exceeds a specified threshold.

Parameters

- **pct_null_threshold** (*float*) – The percentage of NaN values in an input feature to drop. Must be a value between [0, 1] inclusive. If equal to 0.0, will drop columns with any null values. If equal to 1.0, will drop columns with all null values. Defaults to 0.95.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Null Columns Transformer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by dropping columns that exceed the threshold of null values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data *X* by dropping columns that exceed the threshold of null values.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class *evalml.pipelines.components.transformers.preprocessing.DropRowsTransformer* (*indices_to_drop*, *random_seed=0*)

Transformer to drop rows specified by row indices.

Parameters

- **indices_to_drop** (*list*) – List of indices to drop in the input data. Defaults to *None*.
- **random_seed** (*int*) – Seed for the random number generator. Is not used by this component. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Drop Rows Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.preprocessing.**EmailFeaturizer** (*random_seed=0*,
***kwargs*)

Transformer that can automatically extract features from emails.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Email Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self***fit_transform** (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X**Return type** *pd.DataFrame***class** evalml.pipelines.components.transformers.preprocessing.**LogTransformer** (*random_seed=0*)

Applies a log transformation to the target data.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Log Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the LogTransformer.
<i>fit_transform</i>	Log transforms the target variable.
<i>inverse_transform</i>	Inverts the transformation done by the transform method.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Log transforms the target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the LogTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Ignored.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to log transform.

Returns

The input features are returned without modification. The target variable *y* is log transformed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Inverts the transformation done by the transform method.

Arguments: *y* (*pd.Series*): Target transformed by this component.

Returns Target without the transformation.

Return type *pd.Series*∅

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target data to log transform.

Returns

The input features are returned without modification. The target variable `y` is log transformed.

Return type tuple of `pd.DataFrame`, `pd.Series`

class `evalml.pipelines.components.transformers.preprocessing.LSA` (`random_seed=0`,
**`kwargs`)

Transformer to calculate the Latent Semantic Analysis Values of text input.

Parameters `random_seed` (`int`) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	LSA Transformer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on <code>X</code> and transforms <code>X</code>
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data <code>X</code> by applying the LSA pipeline.

clone (`self`)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (`cls`)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (`self`, `print_name=False`, `return_dict=False`)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by applying the LSA pipeline.

Parameters

- **X** (*pd.DataFrame*) – The data to transform.

- `y` (`pd.Series`, *optional*) – Ignored.

Returns

Transformed X. The original column is removed and replaced with two columns of the format `LSA(original_column_name)[feature_number]`, where `feature_number` is 0 or 1.

Return type `pd.DataFrame`

class `evalml.pipelines.components.transformers.preprocessing.PolynomialDetrender` (*degree=1*, *random_seed=0*, ***kwargs*)

Removes trends from time series by fitting a polynomial to the data.

Parameters

- **degree** (*int*) – Degree for the polynomial. If 1, linear model is fit to the data. If 2, quadratic model is fit, etc. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “degree”: Integer(1, 3)}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Polynomial Detrender

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the PolynomialDetrender.
<code>fit_transform</code>	Removes fitted trend from target variable.
<code>inverse_transform</code>	Adds back fitted trend to target variable.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Removes fitted trend from target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the PolynomialDetrender.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns self

fit_transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the fitted trend removed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Adds back fitted trend to target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the trend added back.

Return type tuple of *pd.DataFrame*, *pd.Series*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The input features are returned without modification. The target variable *y* is detrended

Return type tuple of *pd.DataFrame*, *pd.Series*

class evalml.pipelines.components.transformers.preprocessing.**TextFeaturizer** (*random_seed=0*, ***kwargs*)

Transformer that can automatically featurize text columns using featuretools' nlp_primitives.

Since models cannot handle non-numeric data, any text must be broken down into features that provide useful information about that text. This component splits each text column into several informative features: Diversity Score, Mean Characters per Word, Polarity Score, and LSA (Latent Semantic Analysis). Calling transform on this component will replace any text columns in the given dataset with these numeric columns.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Text Featurization Component

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by creating new features using existing text columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by creating new features using existing text columns

Parameters

- **X** (*pd.DataFrame*) – The data to transform.
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.preprocessing.**TextTransformer** (*component_obj=None, random_seed=0, **kwargs*)

Base class for all transformers working with text features.

Parameters

- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>name</code>	Returns string name of this component
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.preprocessing.**URLFeaturizer** (*random_seed=0, **kwargs*)

Transformer that can automatically extract features from URL.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	URL Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

samplers

Submodules

base_sampler

Module Contents

Classes Summary

<i>BaseSampler</i>	Base Sampler component. Used as the base class of all sampler components.
--------------------	---

Contents

class evalml.pipelines.components.transformers.samplers.base_sampler.**BaseSampler** (*parameters=None, component_obj=None, random_seed=0, **kwargs*)

Base Sampler component. Used as the base class of all sampler components.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component

continues on next page

Table 479 – continued from previous page

<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type pd.DataFrame, pd.Series

oversampler

Module Contents

Classes Summary

Oversampler

SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMO-TENC based on inputs to the component.

Contents

class evalml.pipelines.components.transformers.samplers.oversampler.**Oversampler** (*sampling_ratio*, *sampling_ratio_dict*, *k_neighbors_default*, *n_jobs*, *random_seed*, ***kwargs*)

SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.

Parameters

- **sampling_ratio** (*float*) – This is the goal ratio of the minority to majority class, with range (0, 1]. A value of 0.25 means we want a 1:4 ratio of the minority to majority class after oversampling. We will create the a sampling dictionary using this ratio, with the keys corresponding to the class and the values responding to the number of samples. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict*={0: 0.5, 1: 1}, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don't sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **k_neighbors_default** (*int*) – The number of nearest neighbors used to construct synthetic samples. This is the default value used, but the actual *k_neighbors* value might be smaller if there are less samples. Defaults to 5.
- **n_jobs** (*int*) – The number of CPU cores to use. Defaults to -1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Oversampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters

continues on next page

Table 481 – continued from previous page

<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type `pd.DataFrame`, `pd.Series`

undersampler

Module Contents

Classes Summary

Undersampler

Initializes an undersampling transformer to downsample the majority classes in the dataset.

Contents

```
class evalml.pipelines.components.transformers.samplers.undersampler. Undersampler (sampling_ratio, sampling_ratio_min_samples, min_percent, random_seed=0, **kwargs)
```

Initializes an undersampling transformer to downsample the majority classes in the dataset.

This component is only run during training and not during predict.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: `sampling_ratio_dict={0: 0.5, 1: 1}`, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don’t sample class 1. Overrides `sampling_ratio` if provided. Defaults to None.
- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below `min_percentage`. Must be greater than 0. Defaults to 100.
- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above `min_samples`. If `min_percentage` and `min_samples` are not met, treat this as severely imbalanced, and we will not resample the data. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Undersampler

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the sampler to the data.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms the input data by sampling the data.

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type *pd.DataFrame, pd.Series*

Package Contents

Classes Summary

<i>Oversampler</i>	SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.
<i>Undersampler</i>	Initializes an undersampling transformer to downsample the majority classes in the dataset.

Contents

```
class evalml.pipelines.components.transformers.samplers.Oversampler (sampling_ratio=0.25,  
                                                                    sam-  
                                                                    pling_ratio_dict=None,  
                                                                    k_neighbors_default=5,  
                                                                    n_jobs=-  
                                                                    1, ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.

Parameters

- **sampling_ratio** (*float*) – This is the goal ratio of the minority to majority class, with range (0, 1]. A value of 0.25 means we want a 1:4 ratio of the minority to majority class after oversampling. We will create the a sampling dictionary using this ratio, with the keys corresponding to the class and the values responding to the number of samples. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict={0: 0.5, 1: 1}*, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don't sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **k_neighbors_default** (*int*) – The number of nearest neighbors used to construct synthetic samples. This is the default value used, but the actual *k_neighbors* value might be

smaller if there are less samples. Defaults to 5.

- **n_jobs** (*int*) – The number of CPU cores to use. Defaults to -1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Oversampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns *self*

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type *pd.DataFrame*, *pd.Series*


```
class evalml.pipelines.components.transformers.samplers.Undersampler(sampling_ratio=0.25,
                                                                    sam-
                                                                    pling_ratio_dict=None,
                                                                    min_samples=100,
                                                                    min_percentage=0.1,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Initializes an undersampling transformer to downsample the majority classes in the dataset.

This component is only run during training and not during predict.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict={0: 0.5, 1: 1}*, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don’t sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below *min_percentage*. Must be greater than 0. Defaults to 100.
- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above *min_samples*. If *min_percentage* and *min_samples* are not met, treat this as severely imbalanced, and we will not resample the data. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Undersampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X

continues on next page

Table 486 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type pd.DataFrame, pd.Series

scalers

Submodules

standard_scaler

Module Contents

Classes Summary

StandardScaler

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Contents

class evalml.pipelines.components.transformers.scalers.standard_scaler.**StandardScaler** (*random_state=None*, ***kwargs*)

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Standard Scaler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

Package Contents

Classes Summary

<i>StandardScaler</i>	A transformer that standardizes input features by removing the mean and scaling to unit variance.
-----------------------	---

Contents

class evalml.pipelines.components.transformers.scalers.**StandardScaler** (*random_seed=0*,
***kwargs*)

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Standard Scaler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)
Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

Submodules

column_selectors

Module Contents

Classes Summary

<i>ColumnSelector</i>	Initializes an transformer that drops specified columns in input data.
<i>DropColumns</i>	Drops specified columns in input data.
<i>SelectByType</i>	Selects columns by specified Woodwork logical type or semantic tag in input data.
<i>SelectColumns</i>	Selects specified columns in input data.

Contents

class evalml.pipelines.components.transformers.column_selectors.**ColumnSelector** (*columns=None*, *random_seed=0*, ***kwargs*)

Initializes an transformer that drops specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.

- **y** (*pd.Series, optional*) – Targets.

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

class evalml.pipelines.components.transformers.column_selectors.**DropColumns** (*columns=None, random_seed=0, **kwargs*)

Drops specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to drop.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Columns Transformer
needs_fitting	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by dropping columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by dropping columns.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Targets.

Returns Transformed X.

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.column_selectors.**SelectByType** (*column_types=None*,
random_seed=0,
***kwargs*)

Selects columns by specified Woodwork logical type or semantic tag in input data.

Parameters

- **column_types** (*string, ww.LogicalType, list(string), list(ww.LogicalType)*) – List of Woodwork types or tags, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Select Columns By Type Transformer
needs_fitting	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self, print_name=False, return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.column_selectors.**SelectColumns** (*columns=None*,
ran-
dom_seed=0,
***kwargs*)

Selects specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Select Columns Transformer
needs_fitting	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by selecting columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **x** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by selecting columns.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Targets.

Returns Transformed X.

Return type *pd.DataFrame*

transformer

Module Contents

Classes Summary

<i>TargetTransformer</i>	A component that transforms the target.
<i>Transformer</i>	A component that may or may not need fitting that transforms data.

Contents

class evalml.pipelines.components.transformers.transformer.**TargetTransformer** (*parameters=None, component_obj=None, random_seed=0, **kwargs*)

A component that transforms the target.

Attributes

model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>inverse_transform</i>	Inverts the transformation done by the transform method.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

abstract inverse_transform (*self*, *y*)

Inverts the transformation done by the transform method.

Arguments: *y* (*pd.Series*): Target transformed by this component.

Returns Target without the transformation.

Return type pd.Series

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.components.transformers.transformer.Transformer (parameters=None,
                                                                    com-
                                                                    po-
                                                                    nent_obj=None,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

A component that may or may not need fitting that transforms data. These components are used before an estimator.

To implement a new Transformer, define your own class which is a subclass of Transformer, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Transformer component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]

- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

Package Contents

Classes Summary

<i>DateTimeFeaturizer</i>	Transformer that can automatically extract features from datetime columns.
<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
<i>DropColumns</i>	Drops specified columns in input data.
<i>DropNullColumns</i>	Transformer to drop features whose percentage of NaN values exceeds a specified threshold.
<i>DropRowsTransformer</i>	Transformer to drop rows specified by row indices.
<i>EmailFeaturizer</i>	Transformer that can automatically extract features from emails.
<i>FeatureSelector</i>	Selects top features based on importance weights.
<i>Imputer</i>	Imputes missing data according to a specified imputation strategy.
<i>LinearDiscriminantAnalysis</i>	Reduces the number of features by using Linear Discriminant Analysis.
<i>LogTransformer</i>	Applies a log transformation to the target data.
<i>LSA</i>	Transformer to calculate the Latent Semantic Analysis Values of text input.
<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>Oversampler</i>	SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.
<i>PCA</i>	Reduces the number of features by using Principal Component Analysis (PCA).
<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
<i>PolynomialDetrender</i>	Removes trends from time series by fitting a polynomial to the data.
<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.
<i>SelectByType</i>	Selects columns by specified Woodwork logical type or semantic tag in input data.
<i>SelectColumns</i>	Selects specified columns in input data.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>StandardScaler</i>	A transformer that standardizes input features by removing the mean and scaling to unit variance.
<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.
<i>TargetImputer</i>	Imputes missing target data according to a specified imputation strategy.

continues on next page

Table 499 – continued from previous page

<i>TextFeaturizer</i>	Transformer that can automatically featurize text columns using featuretools’ nlp_primitives.
<i>Transformer</i>	A component that may or may not need fitting that transforms data.
<i>Undersampler</i>	Initializes an undersampling transformer to downsample the majority classes in the dataset.
<i>URLFeaturizer</i>	Transformer that can automatically extract features from URL.

Contents

class evalml.pipelines.components.transformers.**DateTimeFeaturizer** (*features_to_extract=None, encode_as_categories=False, date_index=None, random_seed=0, **kwargs*)

Transformer that can automatically extract features from datetime columns.

Parameters

- **features_to_extract** (*list*) – List of features to extract. Valid options include “year”, “month”, “day_of_week”, “hour”. Defaults to None.
- **encode_as_categories** (*bool*) – Whether day-of-week and month features should be encoded as pandas “category” dtype. This allows OneHotEncoders to encode these features. Defaults to False.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DateTime Featurization Component

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data

continues on next page

Table 500 – continued from previous page

<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Gets the categories of each datetime feature.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

get_feature_names (*self*)

Gets the categories of each datetime feature.

Returns Dictionary, where each key-value pair is a column name and a dictionary mapping the unique feature values to their integer encoding.

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.components.transformers.DelayedFeatureTransformer (date_index=None,
                                                                    max_delay=2,
                                                                    gap=0,
                                                                    fore-
                                                                    cast_horizon=1,
                                                                    de-
                                                                    lay_features=True,
                                                                    de-
                                                                    lay_target=True,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Transformer that delays input features and target variable for time series problems.

Parameters

- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **max_delay** (*int*) – Maximum number of time units to delay each feature. Defaults to 2.
- **forecast_horizon** (*int*) – The number of time periods the pipeline is expected to forecast.
- **delay_features** (*bool*) – Whether to delay the input features. Defaults to True.
- **delay_target** (*bool*) – Whether to delay the target. Defaults to True.
- **gap** (*int*) – The number of time units between when the features are collected and when the target is collected. For example, if you are predicting the next time step’s target, gap=1. This is only needed because when gap=0, we need to be sure to start the lagging of the target variable at 1. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. This transformer performs the same regardless of the random seed provided.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Delayed Feature Transformer
needs_fitting	False
target_colname_prefix	target_delay_{ }

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DelayFeatureTransformer.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the delayed features for all features in X and y.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DelayFeatureTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Computes the delayed features for all features in *X* and *y*.

For each feature in *X*, it will add a column to the output dataframe for each delay in the (inclusive) range [1, *max_delay*]. The values of each delayed feature are simply the original feature shifted forward in time by the delay amount. For example, a delay of 3 units means that the feature value at row *n* will be taken from the *n*-3rd row of that feature

If *y* is not *None*, it will also compute the delayed values for the target variable.

Parameters

- **X** (*pd.DataFrame* or *None*) – Data to transform. *None* is expected when only the target variable is being used.
- **y** (*pd.Series*, or *None*) – Target.

Returns Transformed *X*.

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.DFSTransformer (*index='index'*,
random_seed=0,
***kwargs*)

Featuretools DFS component that generates features for the input features.

Parameters

- **index** (*string*) – The name of the column that contains the indices. If no column with this name exists, then `featuretools.EntitySet()` creates a column with this name to serve as the index column. Defaults to 'index'.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DFSTransformer Transformer component.
<i>fit_transform</i>	Fits on <i>X</i> and transforms <i>X</i>
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 502 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the feature matrix for the input X using featuretools' dfs algorithm.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DFSTransformer Transformer component.

Parameters

- **X** (*pd.DataFrame*, *np.array*) – The input data to transform, of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

```
needs_fitting(self)
```

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

```
save (self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL)
```

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Computes the feature matrix for the input X using featuretools' dfs algorithm.

Parameters

- **x** (*pd.DataFrame* or *np.ndarray*) – The input training data to transform. Has shape [n_samples, n_features]
- **y** (*pd.Series*, optional) – Ignored.

Returns Feature matrix

Return type `pd.DataFrame`

[illegible]

Drops specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to drop.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Columns Transformer
needs_fitting	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the transformer by checking if column names are present in the dataset.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by dropping columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns `ComponentBase` object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- `file_path` (*str*) – Location to save file
- `pickle_protocol` (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Transforms data *X* by dropping columns.

Parameters

- `X` (`pd.DataFrame`) – Data to transform.
- `y` (`pd.Series`, *optional*) – Targets.

Returns Transformed *X*.

Return type `pd.DataFrame`

class `evalml.pipelines.components.transformers.DropNullColumns` (*pct_null_threshold*=1.0,
random_seed=0,
***kwargs*)

Transformer to drop features whose percentage of NaN values exceeds a specified threshold.

Parameters

- `pct_null_threshold` (*float*) – The percentage of NaN values in an input feature to drop. Must be a value between [0, 1] inclusive. If equal to 0.0, will drop columns with any null values. If equal to 1.0, will drop columns with all null values. Defaults to 0.95.
- `random_seed` (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
name	Drop Null Columns Transformer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by dropping columns that exceed the threshold of null values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data *X* by dropping columns that exceed the threshold of null values.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class *evalml.pipelines.components.transformers.DropRowsTransformer* (*indices_to_drop=None*,
random_seed=0)

Transformer to drop rows specified by row indices.

Parameters

- **indices_to_drop** (*list*) – List of indices to drop in the input data. Defaults to *None*.
- **random_seed** (*int*) – Seed for the random number generator. Is not used by this component. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Drop Rows Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.**EmailFeaturizer** (*random_seed=0*,
***kwargs*)

Transformer that can automatically extract features from emails.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Email Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self***fit_transform** (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X**Return type** *pd.DataFrame*

```
class evalml.pipelines.components.transformers.FeatureSelector (parameters=None,  

component_obj=None,  

random_seed=0,  

**kwargs)
```

Selects top features based on importance weights.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.Imputer (categorical_impute_strategy='most_frequent',
                                                    categori-
                                                    cal_fill_value=None, nu-
                                                    meric_impute_strategy='mean',
                                                    numeric_fill_value=None,
                                                    random_seed=0, **kwargs)
```

Imputes missing data according to a specified imputation strategy.

Parameters

- **categorical_impute_strategy** (*string*) – Impute strategy to use for string, object, boolean, categorical dtypes. Valid values include “most_frequent” and “constant”.
- **numeric_impute_strategy** (*string*) – Impute strategy to use for numeric columns. Valid values include “mean”, “median”, “most_frequent”, and “constant”.
- **categorical_fill_value** (*string*) – When `categorical_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of None will fill with the string “missing_value”.
- **numeric_fill_value** (*int*, *float*) – When `numeric_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of None will fill with 0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “categorical_impute_strategy”: [“most_frequent”], “numeric_impute_strategy”: [“mean”, “median”, “most_frequent”], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Imputer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits imputer to data. 'None' values are converted to np.nan before imputation and are
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by imputing missing values. 'None' values are converted to np.nan before imputation and are

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. 'None' values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on *X* and transforms *X*

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed *X*

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data *X* by imputing missing values. ‘None’ values are converted to *np.nan* before imputation and are treated as the same.

Parameters

- **x** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed *X*

Return type *pd.DataFrame*

class *evalml.pipelines.components.transformers.LinearDiscriminantAnalysis* (*n_components=None*,
random_seed=0,
***kwargs*)

Reduces the number of features by using Linear Discriminant Analysis.

Parameters

- **n_components** (*int*) – The number of features to maintain after computation. Defaults to *None*.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Linear Discriminant Analysis Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self***fit_transform** (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X**Return type** *pd.DataFrame***class** evalml.pipelines.components.transformers.**LogTransformer** (*random_seed=0*)

Applies a log transformation to the target data.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Log Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the LogTransformer.
<i>fit_transform</i>	Log transforms the target variable.
<i>inverse_transform</i>	Inverts the transformation done by the transform method.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Log transforms the target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the LogTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Ignored.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to log transform.

Returns

The input features are returned without modification. The target variable *y* is log transformed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Inverts the transformation done by the transform method.

Arguments: *y* (*pd.Series*): Target transformed by this component.

Returns Target without the transformation.

Return type *pd.Series*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target data to log transform.

Returns

The input features are returned without modification. The target variable `y` is log transformed.

Return type tuple of `pd.DataFrame`, `pd.Series`

class `evalml.pipelines.components.transformers.LSA` (*random_seed=0, **kwargs*)

Transformer to calculate the Latent Semantic Analysis Values of text input.

Parameters `random_seed` (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	LSA Transformer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on <code>X</code> and transforms <code>X</code>
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data <code>X</code> by applying the LSA pipeline.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by applying the LSA pipeline.

Parameters

- **X** (*pd.DataFrame*) – The data to transform.

- **y** (*pd.Series*, *optional*) – Ignored.

Returns

Transformed X. The original column is removed and replaced with two columns of the format *LSA(original_column_name)[feature_number]*, where *feature_number* is 0 or 1.

Return type `pd.DataFrame`

```
class evalml.pipelines.components.transformers.OneHotEncoder (top_n=10, features_to_encode=None, categories=None, drop='if_binary', handle_unknown='ignore', handle_missing='error', random_seed=0, **kwargs)
```

A transformer that encodes categorical features in a one-hot numeric array.

Parameters

- **top_n** (*int*) – Number of categories per column to encode. If *None*, all categories will be encoded. Otherwise, the *n* most frequent will be encoded and all others will be dropped. Defaults to 10.
- **features_to_encode** (*list[str]*) – List of columns to encode. All other columns will remain untouched. If *None*, all appropriate columns will be encoded. Defaults to *None*.
- **categories** (*list*) – A two dimensional list of categories, where *categories[i]* is a list of the categories for the column at index *i*. This can also be *None*, or “*auto*” if *top_n* is not *None*. Defaults to *None*.
- **drop** (*string*, *list*) – Method (“first” or “if_binary”) to use to drop one category per feature. Can also be a list specifying which categories to drop for each feature. Defaults to “if_binary”.
- **handle_unknown** (*string*) – Whether to ignore or error for unknown categories for a feature encountered during *fit* or *transform*. If either *top_n* or *categories* is used to limit the number of categories per column, this must be “ignore”. Defaults to “ignore”.
- **handle_missing** (*string*) – Options for how to handle missing (NaN) values encountered during *fit* or *transform*. If this is set to “as_category” and NaN values are within the *n* most frequent, “nan” values will be encoded as their own column. If this is set to “error”, any missing values encountered will raise an error. Defaults to “error”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	One Hot Encoder

Methods

<code>categories</code>	Returns a list of the unique categories to be encoded for the particular feature, in order.
<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the categorical features after fitting.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	One-hot encode the input data.

categories (*self*, *feature_name*)

Returns a list of the unique categories to be encoded for the particular feature, in order.

Parameters **feature_name** (*str*) – the name of any feature provided to one-hot encoder during fit

Returns the unique categories, in the same dtype as they were provided during fit

Return type np.ndarray

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self**fit_transform** (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***get_feature_names** (*self*)

Return feature names for the categorical features after fitting.

Feature names are formatted as {column name}_{category name}. In the event of a duplicate name, an integer will be added at the end of the feature name to distinguish it.

For example, consider a dataframe with a column called “A” and category “x_y” and another column called “A_x” with “y”. In this example, the feature names would be “A_x_y” and “A_x_y_1”.

Returns The feature names after encoding, provided in the same order as input_features.**Return type** *np.ndarray***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

One-hot encode the input data.

Parameters

- **x** (*pd.DataFrame*) – Features to one-hot encode.
- **y** (*pd.Series*) – Ignored.

Returns Transformed data, where each categorical feature has been encoded into numerical columns using one-hot encoding.

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.Oversampler(sampling_ratio=0.25,
                                                         sam-
                                                         pling_ratio_dict=None,
                                                         k_neighbors_default=5,
                                                         n_jobs=- 1, ran-
                                                         dom_seed=0,
                                                         **kwargs)
```

SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.

Parameters

- **sampling_ratio** (*float*) – This is the goal ratio of the minority to majority class, with range (0, 1]. A value of 0.25 means we want a 1:4 ratio of the minority to majority class after oversampling. We will create the a sampling dictionary using this ratio, with the keys corresponding to the class and the values responding to the number of samples. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict={0: 0.5, 1: 1}*, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don't sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **k_neighbors_default** (*int*) – The number of nearest neighbors used to construct synthetic samples. This is the default value used, but the actual *k_neighbors* value might be smaller if there are less samples. Defaults to 5.
- **n_jobs** (*int*) – The number of CPU cores to use. Defaults to -1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Oversampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.

continues on next page

Table 513 – continued from previous page

<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

hyper-parameter_ranges	Real(0.25, 1)}:type: {"variance"}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	PCA Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.components.transformers.PerColumnImputer (impute_strategies=None,
                                                                de-
                                                                fault_impute_strategy='most_frequent',
                                                                ran-
                                                                dom_seed=0,
                                                                **kwargs)
```

Imputes missing data according to a specified imputation strategy per column.

Parameters

- **impute_strategies** (*dict*) – Column and {"impute_strategy": strategy, "fill_value":value} pairings. Valid values for impute strategy include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to None, which uses "most_frequent" for all columns. When impute_strategy == "constant", fill_value is used to replace missing data. When None, uses 0 when imputing numerical data and "missing_value" for strings or object data types.
- **default_impute_strategy** (*str*) – Impute strategy to fall back on when none is provided for a certain column. Valid values include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to "most_frequent".
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Per Column Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputers on input data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by imputing missing values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputers on input data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to fit.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by imputing missing values.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to transform.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns Transformed X**Return type** *pd.DataFrame*

```
class evalml.pipelines.components.transformers.PolynomialDetrender (degree=1,  
                                                                ran-  
                                                                dom_seed=0,  
                                                                **kwargs)
```

Removes trends from time series by fitting a polynomial to the data.

Parameters

- **degree** (*int*) – Degree for the polynomial. If 1, linear model is fit to the data. If 2, quadratic model is fit, etc. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "degree": Integer(1, 3)}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Polynomial Detrender

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the PolynomialDetrender.
<i>fit_transform</i>	Removes fitted trend from target variable.
<i>inverse_transform</i>	Adds back fitted trend to target variable.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Removes fitted trend from target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the PolynomialDetrender.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns self

fit_transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the fitted trend removed.

Return type tuple of `pd.DataFrame`, `pd.Series`

inverse_transform (*self*, *y*)

Adds back fitted trend to target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the trend added back.

Return type tuple of `pd.DataFrame`, `pd.Series`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The input features are returned without modification. The target variable `y` is detrended

Return type tuple of *pd.DataFrame*, *pd.Series*

```
class evalml.pipelines.components.transformers.RFClassifierSelectFromModel (number_features=None,
n_estimators=10,
max_depth=None,
per-
cent_features=0.5,
threshold=-
np.inf,
n_jobs=-
1,
ran-
dom_seed=0,
**kwargs)
```

Selects top features based on importance weights using a Random Forest classifier.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.

- **percent_features** (*float*) – Percentage of features to use. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to -np.inf.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Classifier Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X**Return type** `pd.DataFrame`

```
class evalml.pipelines.components.transformers.RFRegressorSelectFromModel (number_features=None,
                                                                    n_estimators=10,
                                                                    max_depth=None,
                                                                    per-
                                                                    cent_features=0.5,
                                                                    threshold=-
                                                                   
                                                                    np.inf,
                                                                    n_jobs=-
                                                                   
                                                                    1,
                                                                    ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

Selects top features based on importance weights using a Random Forest regressor.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to `-np.inf`.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Regressor Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type *list[str]*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an *MethodPropertyNotFoundError* exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.

- **y** (*pd.Series, optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

class `evalml.pipelines.components.transformers.SelectByType` (*column_types=None, random_seed=0, **kwargs*)

Selects columns by specified Woodwork logical type or semantic tag in input data.

Parameters

- **column_types** (*string, ww.LogicalType, list(string), list(ww.LogicalType)*) – List of Woodwork types or tags, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	<code>{}</code>
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
name	Select Columns By Type Transformer
needs_fitting	<code>False</code>

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the transformer by checking if column names are present in the dataset.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type `pd.DataFrame`

class `evalml.pipelines.components.transformers.SelectColumns` (*columns=None*,
random_seed=0,
***kwargs*)

Selects specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	<code>{}</code>
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
name	Select Columns Transformer
needs_fitting	<code>False</code>

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the transformer by checking if column names are present in the dataset.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by selecting columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type `dict`

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **x** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by selecting columns.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Targets.

Returns Transformed X.

Return type `pd.DataFrame`

```
class evalml.pipelines.components.transformers.SimpleImputer (impute_strategy='most_frequent',  
                                                         fill_value=None,  
                                                         random_seed=0,  
                                                         **kwargs)
```

Imputes missing data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types.
- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Simple Imputer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits imputer to data. ‘None’ values are converted to <code>np.nan</code> before imputation and are
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms input by imputing missing values. ‘None’ and <code>np.nan</code> values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. 'None' values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – the target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.transformers.**StandardScaler** (*random_seed=0*,
***kwargs*)

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper- parame- ter_ranges	{}
model_family	ModelFamily.NONE
modi- fies_features	True
modi- fies_target	False
name	Standard Scaler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed *X*

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.TargetEncoder (cols=None,  
                                                         smoothing=1.0,  
                                                         handle_unknown='value',  
                                                         handle_missing='value',  
                                                         random_seed=0,  
                                                         **kwargs)
```

A transformer that encodes categorical features into target encodings.

Parameters

- **cols** (*list*) – Columns to encode. If *None*, all string columns will be encoded, otherwise only the columns provided will be encoded. Defaults to *None*
- **smoothing** (*float*) – The smoothing factor to apply. The larger this value is, the more influence the expected target value has on the resulting target encodings. Must be strictly larger than 0. Defaults to 1.0
- **handle_unknown** (*string*) – Determines how to handle unknown categories for a feature encountered. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **handle_missing** (*string*) – Determines how to handle missing values encountered during *fit* or *transform*. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Target Encoder

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the input features after fitting.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

get_feature_names (*self*)

Return feature names for the input features after fitting.

Returns The feature names after encoding

Return type *np.array*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.TargetImputer (impute_strategy='most_frequent',  
                                                         fill_value=None,  
                                                         random_seed=0,  
                                                         **kwargs)
```

Imputes missing target data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types. Defaults to “most_frequent”.

- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to `None` which uses 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Target Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to target data. ‘None’ values are converted to <code>np.nan</code> before imputation and are
<i>fit_transform</i>	Fits on and transforms the input target data.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input target data by imputing missing values. ‘None’ and <code>np.nan</code> values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y*)

Fits imputer to target data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]. Ignored.
- **y** (*pd.Series, optional*) – The target training data of length [n_samples].

Returns self

fit_transform (*self, X, y*)

Fits on and transforms the input target data.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.
- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (pd.DataFrame, pd.Series)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y*)

Transforms input target data by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.

- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (pd.DataFrame, pd.Series)

class evalml.pipelines.components.transformers.**TextFeaturizer** (*random_seed=0*,
***kwargs*)

Transformer that can automatically featurize text columns using featuretools' nlp_primitives.

Since models cannot handle non-numeric data, any text must be broken down into features that provide useful information about that text. This component splits each text column into several informative features: Diversity Score, Mean Characters per Word, Polarity Score, and LSA (Latent Semantic Analysis). Calling transform on this component will replace any text columns in the given dataset with these numeric columns.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Text Featurization Component

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by creating new features using existing text columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data *X* by creating new features using existing text columns

Parameters

- **X** (*pd.DataFrame*) – The data to transform.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed *X*

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.Transformer (parameters=None,
                                                         component_obj=None,
                                                         random_seed=0,
                                                         **kwargs)
```

A component that may or may not need fitting that transforms data. These components are used before an estimator.

To implement a new Transformer, define your own class which is a subclass of Transformer, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Transformer component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on <i>X</i> and transforms <i>X</i>
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 526 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.transformers.Undersampler (sampling_ratio=0.25,
                                                             sam-
                                                             pling_ratio_dict=None,
                                                             min_samples=100,
                                                             min_percentage=0.1,
                                                             random_seed=0,
                                                             **kwargs)
```

Initializes an undersampling transformer to downsample the majority classes in the dataset.

This component is only run during training and not during predict.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict={0: 0.5, 1: 1}*, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don’t sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below *min_percentage*. Must be greater than 0. Defaults to 100.

- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above min_samples. If min_percentage and min_samples are not met, treat this as severely imbalanced, and we will not resample the data. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Undersampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type *pd.DataFrame*, *pd.Series*

```
class evalml.pipelines.components.transformers.URLFeaturizer(random_seed=0,  
                                                         **kwargs)
```

Transformer that can automatically extract features from URL.

Parameters `random_seed` (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper- parame- ter_ranges	{}
model_family	ModelFamily.NONE
modi- fies_features	True
modi- fies_target	False
name	URL Featurizer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

`default_parameters` (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

`describe` (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

Submodules

component_base

Module Contents

Classes Summary

<i>ComponentBase</i>	Base class for all components.
----------------------	--------------------------------

Contents

class evalml.pipelines.components.component_base.**ComponentBase** (*parameters=None, component_obj=None, random_seed=0, **kwargs*)

Base class for all components.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>model_family</i>	Returns ModelFamily of this component
<i>modifies_features</i>	Returns whether this component modifies (subsets or transforms) the features variable during transform.
<i>modifies_target</i>	Returns whether this component modifies (subsets or transforms) the target variable during transform.
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property model_family (*cls*)

Returns ModelFamily of this component

property modifies_features (*cls*)

Returns whether this component modifies (subsets or transforms) the features variable during transform. For Estimator objects, this attribute determines if the return value from *predict* or *predict_proba* should be used as features or targets.

property modifies_target (*cls*)

Returns whether this component modifies (subsets or transforms) the target variable during transform. For Estimator objects, this attribute determines if the return value from *predict* or *predict_proba* should be used as features or targets.

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

component_base_meta

Module Contents

Classes Summary

<i>ComponentBaseMeta</i>	Metaclass that overrides creating a new component by wrapping methods with validators and setters
--------------------------	---

Contents

class evalml.pipelines.components.component_base_meta.**ComponentBaseMeta**

Metaclass that overrides creating a new component by wrapping methods with validators and setters

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	['predict', 'predict_proba', 'transform', 'inverse_transform']
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<i>check_for_fit</i>	<i>check_for_fit</i> wraps a method that validates if <i>self._is_fitted</i> is <i>True</i> .
<i>register</i>	Register a virtual subclass of an ABC.
<i>set_fit</i>	

classmethod check_for_fit (*cls*, *method*)

check_for_fit wraps a method that validates if *self._is_fitted* is *True*. It raises an exception if *False* and calls and returns the wrapped method if *True*.

register (*cls*, *subclass*)

Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

```
classmethod set_fit (cls, method)
```

utils

Module Contents

Classes Summary

<i>WrappedSKClassifier</i>	Scikit-learn classifier wrapper class.
<i>WrappedSKRegressor</i>	Scikit-learn regressor wrapper class.

Functions

<i>all_components</i>	
<i>allowed_model_families</i>	List the model types allowed for a particular problem type.
<i>generate_component_code</i>	Creates and returns a string that contains the Python imports and code required for running the EvalML component.
<i>get_estimators</i>	Returns the estimators allowed for a particular problem type.
<i>handle_component_class</i>	Standardizes input from a string name to a ComponentBase subclass if necessary.
<i>make_balancing_dictionary</i>	Makes dictionary for oversampler components. Find ratio of each class to the majority.
<i>scikit_learn_wrapped_estimator</i>	

Contents

```
evalml.pipelines.components.utils.all_components()
```

```
evalml.pipelines.components.utils.allowed_model_families(problem_type)
```

List the model types allowed for a particular problem type.

Parameters **problem_types** (*ProblemTypes* or *str*) – binary, multiclass, or regression

Returns a list of model families

Return type list[ModelFamily]

```
evalml.pipelines.components.utils.generate_component_code(element)
```

Creates and returns a string that contains the Python imports and code required for running the EvalML component.

Parameters **element** (*component instance*) – The instance of the component to generate string Python code for

Returns String representation of Python code that can be run separately in order to recreate the component instance. Does not include code for custom component implementation.

`evalml.pipelines.components.utils.get_estimators` (*problem_type*,
model_families=None)

Returns the estimators allowed for a particular problem type.

Can also optionally filter by a list of model types.

Parameters

- **problem_type** (*ProblemTypes* or *str*) – problem type to filter for
- **model_families** (*list[ModelFamily]* or *list[str]*) – model families to filter for

Returns a list of estimator subclasses

Return type `list[class]`

`evalml.pipelines.components.utils.handle_component_class` (*component_class*)

Standardizes input from a string name to a `ComponentBase` subclass if necessary.

If a *str* is provided, will attempt to look up a `ComponentBase` class by that name and return a new instance. Otherwise if a `ComponentBase` subclass or `Component` instance is provided, will return that without modification.

Parameters **component** (*str*, *ComponentBase*) – input to be standardized

Returns `ComponentBase`

`evalml.pipelines.components.utils.make_balancing_dictionary` (*y*, *sampling_ratio*)

Makes dictionary for oversampler components. Find ratio of each class to the majority. If the ratio is smaller than the *sampling_ratio*, we want to oversample, otherwise, we don't want to sample at all, and we leave the data as is.

Parameters

- **y** (*pd.Series*) – Target data
- **sampling_ratio** (*float*) – The balanced ratio we want the samples to meet

Returns Dictionary where keys are the classes, and the corresponding values are the counts of samples for each class that will satisfy *sampling_ratio*.

`evalml.pipelines.components.utils.scikit_learn_wrapped_estimator` (*evalml_obj*)

class `evalml.pipelines.components.utils.WrappedSKClassifier` (*pipeline*)

Scikit-learn classifier wrapper class.

Methods

<i>fit</i>	Fits component to data
<i>get_params</i>	Get parameters for this estimator.
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>score</i>	Return the mean accuracy on the given test data and labels.
<i>set_params</i>	Set the parameters of this estimator.

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape `[n_samples, n_features]`

- **y** (*pd.Series, optional*) – the target training data of length [n_samples]

Returns self

get_params (*self, deep=True*)

Get parameters for this estimator.

Parameters **deep** (*bool, default=True*) – If True, will return the parameters for this estimator and contained subobjects that are estimators.

Returns **params** – Parameter names mapped to their values.

Return type dict

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*) – Features

Returns Predicted values

Return type np.ndarray

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*) – Features

Returns Probability estimates

Return type np.ndarray

score (*self, X, y, sample_weight=None*)

Return the mean accuracy on the given test data and labels.

In multi-label classification, this is the subset accuracy which is a harsh metric since you require for each sample that each label set be correctly predicted.

Parameters

- **X** (*array-like of shape (n_samples, n_features)*) – Test samples.
- **y** (*array-like of shape (n_samples,) or (n_samples, n_outputs)*) – True labels for X.
- **sample_weight** (*array-like of shape (n_samples,)*, *default=None*) – Sample weights.

Returns **score** – Mean accuracy of self.predict(X) wrt. y.

Return type float

set_params (*self, **params*)

Set the parameters of this estimator.

The method works on simple estimators as well as on nested objects (such as Pipeline). The latter have parameters of the form <component>__<parameter> so that it's possible to update each component of a nested object.

Parameters ****params** (*dict*) – Estimator parameters.

Returns **self** – Estimator instance.

Return type estimator instance

class evalml.pipelines.components.utils.**WrappedSKRegressor** (*pipeline*)
Scikit-learn regressor wrapper class.

Methods

<i>fit</i>	Fits component to data
<i>get_params</i>	Get parameters for this estimator.
<i>predict</i>	Make predictions using selected features.
<i>score</i>	Return the coefficient of determination R^2 of the
<i>set_params</i>	Set the parameters of this estimator.

fit (*self*, *X*, *y*)
Fits component to data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [*n_samples*, *n_features*]
- **y** (*pd.Series*, *optional*) – the target training data of length [*n_samples*]

Returns *self*

get_params (*self*, *deep=True*)
Get parameters for this estimator.

Parameters *deep* (*bool*, *default=True*) – If True, will return the parameters for this estimator and contained subobjects that are estimators.

Returns *params* – Parameter names mapped to their values.

Return type *dict*

predict (*self*, *X*)
Make predictions using selected features.

Parameters **X** (*pd.DataFrame*) – Features

Returns Predicted values

Return type *np.ndarray*

score (*self*, *X*, *y*, *sample_weight=None*)
Return the coefficient of determination R^2 of the prediction.

The coefficient R^2 is defined as $(1 - \frac{u}{v})$, where u is the residual sum of squares $((y_{\text{true}} - y_{\text{pred}}) ** 2).sum()$ and v is the total sum of squares $((y_{\text{true}} - y_{\text{true}.mean()}) ** 2).sum()$. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y , disregarding the input features, would get a R^2 score of 0.0.

Parameters

- **X** (*array-like of shape (n_samples, n_features)*) – Test samples. For some estimators this may be a precomputed kernel matrix or a list of generic objects instead with shape $(n_{\text{samples}}, n_{\text{samples_fitted}})$, where *n_samples_fitted* is the number of samples used in the fitting for the estimator.
- **y** (*array-like of shape (n_samples,) or (n_samples, n_outputs)*) – True values for *X*.
- **sample_weight** (*array-like of shape (n_samples,)*, *default=None*) – Sample weights.

Returns `score` – R^2 of `self.predict(X)` wrt. `y`.

Return type float

Notes

The R^2 score used when calling `score` on a regressor uses `multioutput='uniform_average'` from version 0.23 to keep consistent with default value of `r2_score()`. This influences the `score` method of all the multioutput regressors (except for `MultiOutputRegressor`).

set_params (*self*, ****params**)

Set the parameters of this estimator.

The method works on simple estimators as well as on nested objects (such as `Pipeline`). The latter have parameters of the form `<component>__<parameter>` so that it's possible to update each component of a nested object.

Parameters ****params** (*dict*) – Estimator parameters.

Returns `self` – Estimator instance.

Return type estimator instance

Package Contents

Classes Summary

<i>ARIMAREgressor</i>	Autoregressive Integrated Moving Average Model.
<i>BaselineClassifier</i>	Classifier that predicts using the specified strategy.
<i>BaselineRegressor</i>	Baseline regressor that uses a simple strategy to make predictions.
<i>CatBoostClassifier</i>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.
<i>CatBoostRegressor</i>	CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.
<i>ComponentBase</i>	Base class for all components.
<i>ComponentBaseMeta</i>	Metaclass that overrides creating a new component by wrapping methods with validators and setters
<i>DateTimeFeaturizer</i>	Transformer that can automatically extract features from datetime columns.
<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
<i>DropColumns</i>	Drops specified columns in input data.
<i>DropNullColumns</i>	Transformer to drop features whose percentage of NaN values exceeds a specified threshold.
<i>DropRowsTransformer</i>	Transformer to drop rows specified by row indices.
<i>ElasticNetClassifier</i>	Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.
<i>ElasticNetRegressor</i>	Elastic Net Regressor.

continues on next page

Table 537 – continued from previous page

<i>EmailFeaturizer</i>	Transformer that can automatically extract features from emails.
<i>Estimator</i>	A component that fits and predicts given data.
<i>ExtraTreesClassifier</i>	Extra Trees Classifier.
<i>ExtraTreesRegressor</i>	Extra Trees Regressor.
<i>FeatureSelector</i>	Selects top features based on importance weights.
<i>Imputer</i>	Imputes missing data according to a specified imputation strategy.
<i>KNeighborsClassifier</i>	K-Nearest Neighbors Classifier.
<i>LightGBMClassifier</i>	LightGBM Classifier.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>LinearDiscriminantAnalysis</i>	Reduces the number of features by using Linear Discriminant Analysis.
<i>LinearRegressor</i>	Linear Regressor.
<i>LogisticRegressionClassifier</i>	Logistic Regression Classifier.
<i>LogTransformer</i>	Applies a log transformation to the target data.
<i>LSA</i>	Transformer to calculate the Latent Semantic Analysis Values of text input.
<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>Oversampler</i>	SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMO-TENC based on inputs to the component.
<i>PCA</i>	Reduces the number of features by using Principal Component Analysis (PCA).
<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
<i>PolynomialDetrender</i>	Removes trends from time series by fitting a polynomial to the data.
<i>ProphetRegressor</i>	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
<i>RandomForestClassifier</i>	Random Forest Classifier.
<i>RandomForestRegressor</i>	Random Forest Regressor.
<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.
<i>SelectByType</i>	Selects columns by specified Woodwork logical type or semantic tag in input data.
<i>SelectColumns</i>	Selects specified columns in input data.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>SklearnStackedEnsembleClassifier</i>	Scikit-learn Stacked Ensemble Classifier.
<i>SklearnStackedEnsembleRegressor</i>	Scikit-learn Stacked Ensemble Regressor.
<i>StandardScaler</i>	A transformer that standardizes input features by removing the mean and scaling to unit variance.
<i>SVMClassifier</i>	Support Vector Machine Classifier.
<i>SVMRegressor</i>	Support Vector Machine Regressor.

continues on next page

Table 537 – continued from previous page

<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.
<i>TargetImputer</i>	Imputes missing target data according to a specified imputation strategy.
<i>TextFeaturizer</i>	Transformer that can automatically featurize text columns using featuretools' nlp_primitives.
<i>TimeSeriesBaselineEstimator</i>	Time series estimator that predicts using the naive forecasting approach.
<i>Transformer</i>	A component that may or may not need fitting that transforms data.
<i>Undersampler</i>	Initializes an undersampling transformer to downsample the majority classes in the dataset.
<i>URLFeaturizer</i>	Transformer that can automatically extract features from URL.
<i>XGBoostClassifier</i>	XGBoost Classifier.
<i>XGBoostRegressor</i>	XGBoost Regressor.

Contents

```
class evalml.pipelines.components.ARIMAREgressor (date_index=None, trend=None,
                                                    start_p=2, d=0, start_q=2,
                                                    max_p=5, max_d=2, max_q=5,
                                                    seasonal=True, n_jobs=-1, random_seed=0, **kwargs)
```

Autoregressive Integrated Moving Average Model. The three parameters (p, d, q) are the AR order, the degree of differencing, and the MA order. More information here: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

Currently ARIMAREgressor isn't supported via conda install. It's recommended that it be installed via PyPI.

Parameters

- **date_index** (*str*) – Specifies the name of the column in X that provides the datetime objects. Defaults to None.
- **trend** (*str*) – Controls the deterministic trend. Options are ['n', 'c', 't', 'ct'] where 'c' is a constant term, 't' indicates a linear trend, and 'ct' is both. Can also be an iterable when defining a polynomial, such as [1, 1, 0, 1].
- **start_p** (*int*) – Minimum Autoregressive order. Defaults to 2.
- **d** (*int*) – Minimum Differencing degree. Defaults to 0.
- **start_q** (*int*) – Minimum Moving Average order. Defaults to 2.
- **max_p** (*int*) – Maximum Autoregressive order. Defaults to 5.
- **max_d** (*int*) – Maximum Differencing degree. Defaults to 2.
- **max_q** (*int*) – Maximum Moving Average order. Defaults to 5.
- **seasonal** (*boolean*) – Whether to fit a seasonal model to ARIMA. Defaults to True.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "start_p": Integer(1, 3), "d": Integer(0, 2), "start_q": Integer(1, 3), "max_p": Integer(3, 10), "max_d": Integer(2, 5), "max_q": Integer(3, 10), "seasonal": [True, False], }
model_family	ModelFamily.ARIMA
modifies_features	True
modifies_target	False
name	ARIMA Regressor
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*, *y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**BaselineClassifier** (*strategy='mode'*, *random_seed=0*, ***kwargs*)

Classifier that predicts using the specified strategy.

This is useful as a simple baseline classifier to compare with other classifiers.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mode”, “random” and “random_weighted”. Defaults to “mode”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS]

Methods

<i>classes_</i>	Returns class labels. Will return None before fitting.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

property classes_ (*self*)

Returns class labels. Will return None before fitting.

Returns Class names

Return type list[str] or list(float)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**BaselineRegressor** (*strategy='mean'*, *random_seed=0*, ***kwargs*)

Baseline regressor that uses a simple strategy to make predictions. This is useful as a simple baseline regressor to compare with other regressors.

Parameters

- **strategy** (*str*) – Method used to predict. Valid options are “mean”, “median”. Defaults to “mean”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Baseline Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.
<i>fit</i>	Fits component to data

continues on next page

Table 540 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature. Since baseline regressors do not use input features to calculate predictions, returns an array of zeroes.

Returns An array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.CatBoostClassifier (n_estimators=10, eta=0.03,  
                                                    max_depth=6, bootstrap_type=None, silent=True,  
                                                    allow_writing_files=False,  
                                                    random_seed=0, n_jobs=-1,  
                                                    **kwargs)
```

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.

- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.CatBoostRegressor (n_estimators=10, eta=0.03,  
                                                    max_depth=6, bootstrap_type=None,  
                                                    silent=False,  
                                                    allow_writing_files=False,  
                                                    random_seed=0, n_jobs=-1,  
                                                    **kwargs)
```

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters

continues on next page

Table 542 – continued from previous page

<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**ComponentBase** (*parameters=None*, *component_obj=None*, *random_seed=0*, ***kwargs*)

Base class for all components.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>model_family</i>	Returns ModelFamily of this component

continues on next page

Table 543 – continued from previous page

<i>modifies_features</i>	Returns whether this component modifies (subsets or transforms) the features variable during transform.
<i>modifies_target</i>	Returns whether this component modifies (subsets or transforms) the target variable during transform.
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property model_family (*cls*)

Returns ModelFamily of this component

property modifies_features (*cls*)

Returns whether this component modifies (subsets or transforms) the features variable during transform. For Estimator objects, this attribute determines if the return value from *predict* or *predict_proba* should be used as features or targets.

property modifies_target (*cls*)

Returns whether this component modifies (subsets or transforms) the target variable during transform. For Estimator objects, this attribute determines if the return value from *predict* or *predict_proba* should be used as features or targets.

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**ComponentBaseMeta**

Metaclass that overrides creating a new component by wrapping methods with validators and setters

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	['predict', 'predict_proba', 'transform', 'inverse_transform']
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<i>check_for_fit</i>	<i>check_for_fit</i> wraps a method that validates if <i>self.is_fitted</i> is <i>True</i> .
<i>register</i>	Register a virtual subclass of an ABC.
<i>set_fit</i>	

classmethod **check_for_fit** (*cls*, *method*)

check_for_fit wraps a method that validates if *self.is_fitted* is *True*. It raises an exception if *False* and calls and returns the wrapped method if *True*.

register (*cls*, *subclass*)

Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

```
classmethod set_fit(cls, method)
```

```
class evalml.pipelines.components.DateTimeFeaturizer(features_to_extract=None, encode_as_categories=False, date_index=None, random_seed=0, **kwargs)
```

Transformer that can automatically extract features from datetime columns.

Parameters

- **features_to_extract** (*list*) – List of features to extract. Valid options include “year”, “month”, “day_of_week”, “hour”. Defaults to None.
- **encode_as_categories** (*bool*) – Whether day-of-week and month features should be encoded as pandas “category” dtype. This allows OneHotEncoders to encode these features. Defaults to False.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DateTime Featurization Component

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Gets the categories of each datetime feature.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by creating new features using existing DateTime columns, and then dropping those DateTime columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Gets the categories of each datetime feature.

Returns Dictionary, where each key-value pair is a column name and a dictionary mapping the unique feature values to their integer encoding.

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Transforms data *X* by creating new features using existing `DateTime` columns, and then dropping those `DateTime` columns

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed *X*

Return type `pd.DataFrame`

```
class evalml.pipelines.components.DecisionTreeClassifier (criterion='gini',  
                                                         max_features='auto',  
                                                         max_depth=6,  
                                                         min_samples_split=2,  
                                                         min_weight_fraction_leaf=0.0,  
                                                         random_seed=0,  
                                                         **kwargs)
```

Decision Tree Classifier.

Parameters

- **criterion** (*{ "gini", "entropy" }*) – The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Defaults to “gini”.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If `int`, then consider `max_features` features at each split.
 - If `float`, then `max_features` is a fraction and `int(max_features * n_features)` features are considered at each split.
 - If “auto”, then `max_features=sqrt(n_features)`.
 - If “sqrt”, then `max_features=sqrt(n_features)`.
 - If “log2”, then `max_features=log2(n_features)`.
 - If `None`, then `max_features = n_features`.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than `max_features` features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.
 - If float, then min_samples_split is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
 Defaults to 2.
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["gini", "entropy"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.DecisionTreeRegressor (criterion='mse',
                                                    max_features='auto',
                                                    max_depth=6,
                                                    min_samples_split=2,
                                                    min_weight_fraction_leaf=0.0,
                                                    random_seed=0,
                                                    **kwargs)
```

Decision Tree Regressor.

Parameters

- **criterion** (*{ "mse", "friedman_mse", "mae", "poisson" }*) – The function to measure the quality of a split. Supported criteria are:
 - “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node
 - “friedman_mse”, which uses mean squared error with Friedman’s improvement score for potential splits
 - “mae” for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,
 - “poisson” which uses reduction in Poisson deviance to find splits.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and *int(max_features * n_features)* features are considered at each split.
 - If “auto”, then *max_features=sqrt(n_features)*.
 - If “sqrt”, then *max_features=sqrt(n_features)*.
 - If “log2”, then *max_features=log2(n_features)*.
 - If *None*, then *max_features = n_features*.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.

- If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

Defaults to 2.

- **`min_weight_fraction_leaf`** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **`random_seed`** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “criterion”: [“mse”, “friedman_mse”, “mae”], “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

`clone` (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

`default_parameters` (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.DelayedFeatureTransformer (date_index=None,  
                                                         max_delay=2,  
                                                         gap=0, fore-  
                                                         cast_horizon=1, de-  
                                                         lay_features=True,  
                                                         delay_target=True,  
                                                         random_seed=0,  
                                                         **kwargs)
```

Transformer that delays input features and target variable for time series problems.

Parameters

- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **max_delay** (*int*) – Maximum number of time units to delay each feature. Defaults to 2.
- **forecast_horizon** (*int*) – The number of time periods the pipeline is expected to forecast.
- **delay_features** (*bool*) – Whether to delay the input features. Defaults to True.
- **delay_target** (*bool*) – Whether to delay the target. Defaults to True.
- **gap** (*int*) – The number of time units between when the features are collected and when the target is collected. For example, if you are predicting the next time step’s target, gap=1. This is only needed because when gap=0, we need to be sure to start the lagging of the target variable at 1. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. This transformer performs the same regardless of the random seed provided.

Attributes

hyper-param- eter_ranges	{}
model_family	ModelFamily.NONE
modi- fies_features	True
modi- fies_target	False
name	Delayed Feature Transformer
needs_fitting	False
tar- get_colname_prefix	target_delay_{ }

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the DelayFeatureTransformer.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Computes the delayed features for all features in X and y.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DelayFeatureTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns `ComponentBase` object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- `file_path` (*str*) – Location to save file
- `pickle_protocol` (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Computes the delayed features for all features in *X* and *y*.

For each feature in *X*, it will add a column to the output dataframe for each delay in the (inclusive) range [1, `max_delay`]. The values of each delayed feature are simply the original feature shifted forward in time by the delay amount. For example, a delay of 3 units means that the feature value at row *n* will be taken from the *n*-3rd row of that feature

If *y* is not `None`, it will also compute the delayed values for the target variable.

Parameters

- `X` (`pd.DataFrame` or `None`) – Data to transform. `None` is expected when only the target variable is being used.
- `y` (`pd.Series`, or `None`) – Target.

Returns Transformed *X*.

Return type `pd.DataFrame`

class `evalml.pipelines.components.DFSTransformer` (*index*='index', *random_seed*=0, ***kwargs*)

Featuretools DFS component that generates features for the input features.

Parameters

- `index` (*string*) – The name of the column that contains the indices. If no column with this name exists, then `featuretools.EntitySet()` creates a column with this name to serve as the index column. Defaults to 'index'.
- `random_seed` (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DFSTransformer Transformer component.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the feature matrix for the input X using featuretools' dfs algorithm.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DFSTransformer Transformer component.

Parameters

- **X** (*pd.DataFrame, np.array*) – The input data to transform, of shape [n_samples, n_features]
- **y** (*pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self**fit_transform** (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** pd.DataFrame**static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** ComponentBase object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self, X, y=None*)

Computes the feature matrix for the input X using featuretools' dfs algorithm.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – The input training data to transform. Has shape [n_samples, n_features]
- **y** (*pd.Series, optional*) – Ignored.

Returns Feature matrix**Return type** pd.DataFrame

class evalml.pipelines.components.**DropColumns** (*columns=None, random_seed=0, **kwargs*)

Drops specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to drop.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Columns Transformer
needs_fitting	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by dropping columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **x** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X by dropping columns.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Targets.

Returns Transformed X.

Return type *pd.DataFrame*

class evalml.pipelines.components.DropNullColumns (*pct_null_threshold=1.0*, *random_seed=0*, ***kwargs*)

Transformer to drop features whose percentage of NaN values exceeds a specified threshold.

Parameters

- **pct_null_threshold** (*float*) – The percentage of NaN values in an input feature to drop. Must be a value between [0, 1] inclusive. If equal to 0.0, will drop columns with any null values. If equal to 1.0, will drop columns with all null values. Defaults to 0.95.

- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Drop Null Columns Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by dropping columns that exceed the threshold of null values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms data X by dropping columns that exceed the threshold of null values.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series*, *optional*) – Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.**DropRowsTransformer** (*indices_to_drop=None*, *random_seed=0*)

Transformer to drop rows specified by row indices.

Parameters

- **indices_to_drop** (*list*) – List of indices to drop in the input data. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Is not used by this component. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Drop Rows Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.components.ElasticNetClassifier (penalty='elasticnet',  
                                                    C=1.0, l1_ratio=0.15,  
                                                    multi_class='auto',  
                                                    solver='saga', n_jobs=- 1,  
                                                    random_seed=0, **kwargs)
```

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Parameters

- **penalty** (`{ "l1", "l2", "elasticnet", "none" }`) – The norm used in penalization. Defaults to “elasticnet”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **multi_class** (`{ "auto", "ovr", "multinomial" }`) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when `solver="liblinear"`. “auto” selects “ovr” if the data is binary, or if `solver="liblinear"`, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (`{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }`) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.
 - “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
 - “liblinear” and “saga” also handle L1 penalty
 - “saga” also supports “elasticnet” penalty
 - “liblinear” does not support setting `penalty='none'`
 Defaults to “saga”.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0.01, 10), “l1_ratio”: Real(0, 1)}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.ElasticNetRegressor (alpha=0.0001, l1_ratio=0.15,
                                                    max_iter=1000, normalize=False,
                                                    random_seed=0,
                                                    **kwargs)
```

Elastic Net Regressor.

Parameters

- **alpha** (*float*) – Constant that multiplies the penalty terms. Defaults to 0.0001.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **max_iter** (*int*) – The maximum number of iterations. Defaults to 1000.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. Defaults to False.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "alpha": Real(0, 1), "l1_ratio": Real(0, 1), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Regressor
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**EmailFeaturizer** (*random_seed=0, **kwargs*)
Transformer that can automatically extract features from emails.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Email Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self, print_name=False, return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.Estimator (parameters=None, component_obj=None,  
                                             random_seed=0, **kwargs)
```

A component that fits and predicts given data.

To implement a new Estimator, define your own class which is a subclass of Estimator, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Estimator component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
pre-dict_uses_y	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path
<i>supported_problem_types</i>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

property supported_problem_types (*cls*)

Problem types this estimator supports

```
class evalml.pipelines.components.ExtraTreesClassifier (n_estimators=100,  
                                                    max_features='auto',  
                                                    max_depth=6,  
                                                    min_samples_split=2,  
                                                    min_weight_fraction_leaf=0.0,  
                                                    n_jobs=-1, ran-  
                                                    dom_seed=0, **kwargs)
```

Extra Trees Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int*, *float* or {"auto", "sqrt", "log2"}) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and *int(max_features * n_features)* features are considered at each split.
 - If “auto”, then *max_features=sqrt(n_features)*.
 - If “sqrt”, then *max_features=sqrt(n_features)*.
 - If “log2”, then *max_features=log2(n_features)*.
 - If *None*, then *max_features = n_features*.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int* or *float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and *ceil(min_samples_split * n_samples)* are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.ExtraTreesRegressor (n_estimators=100,
                                                    max_features='auto',
                                                    max_depth=6,
                                                    min_samples_split=2,
                                                    min_weight_fraction_leaf=0.0,
                                                    n_jobs=-1, random_seed=0,
                                                    **kwargs)
```

Extra Trees Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “sqrt”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “log2”, then $\text{max_features} = \log_2(\text{n_features})$.
 - If *None*, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.FeatureSelector (parameters=None, component_obj=None, random_seed=0, **kwargs)
```

Selects top features based on importance weights.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None**transform** (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X**Return type** `pd.DataFrame`

```
class evalml.pipelines.components.Imputer (categorical_impute_strategy='most_frequent',  
                                             categorical_fill_value=None,           nu-  
                                             numeric_impute_strategy='mean',           nu-  
                                             numeric_fill_value=None,           random_seed=0,  
                                             **kwargs)
```

Imputes missing data according to a specified imputation strategy.

Parameters

- **categorical_impute_strategy** (*string*) – Impute strategy to use for string, object, boolean, categorical dtypes. Valid values include “most_frequent” and “constant”.
- **numeric_impute_strategy** (*string*) – Impute strategy to use for numeric columns. Valid values include “mean”, “median”, “most_frequent”, and “constant”.
- **categorical_fill_value** (*string*) – When `categorical_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of `None` will fill with the string “missing_value”.
- **numeric_fill_value** (*int*, *float*) – When `numeric_impute_strategy == “constant”`, `fill_value` is used to replace missing data. The default value of `None` will fill with 0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “categorical_impute_strategy”: [“most_frequent”], “numeric_impute_strategy”: [“mean”, “median”, “most_frequent”], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Imputer

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X by imputing missing values. ‘None’ values are converted to np.nan before imputation and are

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X**Return type** *pd.DataFrame***static load** (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file**Returns** *ComponentBase* object**needs_fitting** (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`**transform** (*self, X, y=None*)

Transforms data X by imputing missing values. ‘None’ values are converted to `np.nan` before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X**Return type** *pd.DataFrame*

```
class evalml.pipelines.components.KNeighborsClassifier (n_neighbors=5,  
                                                    weights='uniform',    algo-  
                                                    rithm='auto',    leaf_size=30,  
                                                    p=2,          random_seed=0,  
                                                    **kwargs)
```

K-Nearest Neighbors Classifier.

Parameters

- **n_neighbors** (*int*) – Number of neighbors to use by default. Defaults to 5.
- **weights** (*{ 'uniform', 'distance' } or callable*) – Weight function used in prediction. Can be:
 - ‘uniform’ : uniform weights. All points in each neighborhood are weighted equally.

- ‘distance’ : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
- [callable] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Defaults to “uniform”.

- **algorithm** ({‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’}) – Algorithm used to compute the nearest neighbors:
 - ‘ball_tree’ will use BallTree
 - ‘kd_tree’ will use KDTree
 - ‘brute’ will use a brute-force search.

‘auto’ will attempt to decide the most appropriate algorithm based on the values passed to fit method. Defaults to “auto”. Note: fitting on sparse input will override the setting of this parameter, using brute force.
- **leaf_size** (*int*) – Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem. Defaults to 30.
- **p** (*int*) – Power parameter for the Minkowski metric. When $p = 1$, this is equivalent to using manhattan_distance (l1), and euclidean_distance (l2) for $p = 2$. For arbitrary p , minkowski_distance (l_p) is used. Defaults to 2.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_neighbors”: Integer(2, 12), “weights”: [“uniform”, “distance”], “algorithm”: [“auto”, “ball_tree”, “kd_tree”, “brute”], “leaf_size”: Integer(10, 30), “p”: Integer(1, 5), }
model_family	ModelFamily.K_NEIGHBORS
modifies_features	True
modifies_target	False
name	KNN Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0’s matching the input number of features as feature_importance is
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path

continues on next page

Table 561 – continued from previous page

<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's matching the input number of features as `feature_importance` is not defined for KNN classifiers.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.LightGBMClassifier (boosting_type='gbdt',
                                                    learning_rate=0.1,
                                                    n_estimators=100,
                                                    max_depth=0, num_leaves=31,
                                                    min_child_samples=20, bag-
                                                    ging_fraction=0.9, bag-
                                                    ging_freq=0, n_jobs=-1,
                                                    random_seed=0, **kwargs)
```

LightGBM Classifier.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.

- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select `bagging_fraction * 100 %` of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1), }
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Classifier
predict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.LightGBMRegressor (boosting_type='gbdt', learning_rate=0.1, n_estimators=20, max_depth=0, num_leaves=31, min_child_samples=20, bagging_fraction=0.9, bagging_freq=0, n_jobs=-1, random_seed=0, **kwargs)
```

LightGBM Regressor.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select `bagging_fraction * 100 %` of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int* or *None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1),}
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Regressor
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**LinearDiscriminantAnalysis** (*n_components=None*,
random_seed=0,
***kwargs*)

Reduces the number of features by using Linear Discriminant Analysis.

Parameters

- **n_components** (*int*) – The number of features to maintain after computation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Linear Discriminant Analysis Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed *X*

Return type *pd.DataFrame*

class evalml.pipelines.components.**LinearRegressor** (*fit_intercept=True*, *normalize=False*, *n_jobs=-1*, *random_seed=0*, ***kwargs*)

Linear Regressor.

Parameters

- **fit_intercept** (*boolean*) – Whether to calculate the intercept for this model. If set to *False*, no intercept will be used in calculations (i.e. data is expected to be centered). Defaults to *True*.
- **normalize** (*boolean*) – If *True*, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. This parameter is ignored when *fit_intercept* is set to *False*. Defaults to *False*.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “fit_intercept”: [True, False], “normalize”: [True, False]}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Linear Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path

continues on next page

Table 565 – continued from previous page

<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.LogisticRegressionClassifier (penalty='l2',
                                                                C=1.0,
                                                                multi_class='auto',
                                                                solver='lbfgs',
                                                                n_jobs=-1, random_seed=0,
                                                                **kwargs)
```

Logistic Regression Classifier.

Parameters

- **penalty** (*{ "l1", "l2", "elasticnet", "none" }*) – The norm used in penalization. Defaults to “l2”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **multi_class** (*{ "auto", "ovr", "multinomial" }*) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when solver=“liblinear”. “auto” selects “ovr” if the data is binary, or if solver=“liblinear”, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (*{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }*) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.

- “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
- “liblinear” and “saga” also handle L1 penalty
- “saga” also supports “elasticnet” penalty
- “liblinear” does not support setting `penalty='none'`

Defaults to “lbfgs”.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “penalty”: [“l2”], “C”: Real(0.01, 10), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Logistic Regression Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

class `evalml.pipelines.components.LogTransformer` (*random_seed=0*)

Applies a log transformation to the target data.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Log Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the LogTransformer.
<i>fit_transform</i>	Log transforms the target variable.
<i>inverse_transform</i>	Inverts the transformation done by the transform method.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Log transforms the target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the LogTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Ignored.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns self

fit_transform (*self*, *X*, *y=None*)

Log transforms the target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to log transform.

Returns

The input features are returned without modification. The target variable y is log transformed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Inverts the transformation done by the transform method.

Arguments: y (*pd.Series*): Target transformed by this component.

Returns Target without the transformation.

Return type *pd.Series*∅

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=*None*)

Log transforms the target variable.

Parameters

- **x** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target data to log transform.

Returns

The input features are returned without modification. The target variable *y* is log transformed.

Return type tuple of *pd.DataFrame*, *pd.Series*

class evalml.pipelines.components.**LSA** (*random_seed*=0, ***kwargs*)

Transformer to calculate the Latent Semantic Analysis Values of text input.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	LSA Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on <i>X</i> and transforms <i>X</i>
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 568 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by applying the LSA pipeline.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms data X by applying the LSA pipeline.

Parameters

- **X** (*pd.DataFrame*) – The data to transform.
- **y** (*pd.Series*, *optional*) – Ignored.

Returns

Transformed X. The original column is removed and replaced with two columns of the format *LSA(original_column_name)[feature_number]*, where *feature_number* is 0 or 1.

Return type *pd.DataFrame*

```
class evalml.pipelines.components.OneHotEncoder (top_n=10, features_to_encode=None,
                                                categories=None, drop='if_binary',
                                                handle_unknown='ignore', handle_missing='error',
                                                random_seed=0,
                                                **kwargs)
```

A transformer that encodes categorical features in a one-hot numeric array.

Parameters

- **top_n** (*int*) – Number of categories per column to encode. If None, all categories will be encoded. Otherwise, the *n* most frequent will be encoded and all others will be dropped. Defaults to 10.
- **features_to_encode** (*list[str]*) – List of columns to encode. All other columns will remain untouched. If None, all appropriate columns will be encoded. Defaults to None.
- **categories** (*list*) – A two dimensional list of categories, where *categories[i]* is a list of the categories for the column at index *i*. This can also be None, or “auto” if *top_n* is not None. Defaults to None.
- **drop** (*string*, *list*) – Method (“first” or “if_binary”) to use to drop one category per feature. Can also be a list specifying which categories to drop for each feature. Defaults to ‘if_binary’.
- **handle_unknown** (*string*) – Whether to ignore or error for unknown categories for a feature encountered during *fit* or *transform*. If either *top_n* or *categories* is used to limit the number of categories per column, this must be “ignore”. Defaults to “ignore”.

- **handle_missing** (*string*) – Options for how to handle missing (NaN) values encountered during *fit* or *transform*. If this is set to “as_category” and NaN values are within the *n* most frequent, “nan” values will be encoded as their own column. If this is set to “error”, any missing values encountered will raise an error. Defaults to “error”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	One Hot Encoder

Methods

<i>categories</i>	Returns a list of the unique categories to be encoded for the particular feature, in order.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Return feature names for the categorical features after fitting.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	One-hot encode the input data.

categories (*self*, *feature_name*)

Returns a list of the unique categories to be encoded for the particular feature, in order.

Parameters **feature_name** (*str*) – the name of any feature provided to one-hot encoder during fit

Returns the unique categories, in the same dtype as they were provided during fit

Return type np.ndarray

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Return feature names for the categorical features after fitting.

Feature names are formatted as {column name}_{category name}. In the event of a duplicate name, an integer will be added at the end of the feature name to distinguish it.

For example, consider a dataframe with a column called “A” and category “x_y” and another column called “A_x” with “y”. In this example, the feature names would be “A_x_y” and “A_x_y_1”.

Returns The feature names after encoding, provided in the same order as input_features.

Return type np.ndarray

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

One-hot encode the input data.

Parameters

- **X** (*pd.DataFrame*) – Features to one-hot encode.
- **y** (*pd.Series*) – Ignored.

Returns Transformed data, where each categorical feature has been encoded into numerical columns using one-hot encoding.

Return type *pd.DataFrame*

```
class evalml.pipelines.components.Oversampler (sampling_ratio=0.25, sampling_ratio_dict=None,
k_neighbors_default=5, n_jobs=-1,
random_seed=0, **kwargs)
```

SMOTE Oversampler component. Will automatically select whether to use SMOTE, SMOTEN, or SMOTENC based on inputs to the component.

Parameters

- **sampling_ratio** (*float*) – This is the goal ratio of the minority to majority class, with range (0, 1]. A value of 0.25 means we want a 1:4 ratio of the minority to majority class after oversampling. We will create the a sampling dictionary using this ratio, with the keys corresponding to the class and the values responding to the number of samples. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: *sampling_ratio_dict={0: 0.5, 1: 1}*, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don't sample class 1. Overrides *sampling_ratio* if provided. Defaults to None.
- **k_neighbors_default** (*int*) – The number of nearest neighbors used to construct synthetic samples. This is the default value used, but the actual *k_neighbors* value might be smaller if there are less samples. Defaults to 5.
- **n_jobs** (*int*) – The number of CPU cores to use. Defaults to -1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Oversampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self, X, y*)
Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)
Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)
Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)
Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)
Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.
- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type *pd.DataFrame, pd.Series*

class evalml.pipelines.components.**PCA** (*variance=0.95, n_components=None, random_seed=0, **kwargs*)

Reduces the number of features by using Principal Component Analysis (PCA).

Parameters

- **variance** (*float*) – The percentage of the original data variance that should be preserved when reducing the number of features. Defaults to 0.95.

- **n_components** (*int*) – The number of features to maintain after computing SVD. Defaults to None, but will override variance variable if set.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	Real(0.25, 1)}:type: {"variance"
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	PCA Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.PerColumnImputer (impute_strategies=None, default_impute_strategy='most_frequent', random_seed=0, **kwargs)
```

Imputes missing data according to a specified imputation strategy per column.

Parameters

- **impute_strategies** (*dict*) – Column and {"impute_strategy": strategy, "fill_value":value} pairings. Valid values for impute strategy include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to None, which uses "most_frequent" for all columns. When impute_strategy == "constant", fill_value is used to replace missing data. When None, uses 0 when imputing numerical data and "missing_value" for strings or object data types.
- **default_impute_strategy** (*str*) – Impute strategy to fall back on when none is provided for a certain column. Valid values include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to "most_frequent".
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Per Column Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputers on input data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by imputing missing values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputers on input data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to fit.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by imputing missing values.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to transform.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.components.**PolynomialDetrender** (*degree=1*, *random_seed=0*, ***kwargs*)

Removes trends from time series by fitting a polynomial to the data.

Parameters

- **degree** (*int*) – Degree for the polynomial. If 1, linear model is fit to the data. If 2, quadratic model is fit, etc. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “degree”: Integer(1, 3)}
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Polynomial Detrender

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the PolynomialDetrender.
<i>fit_transform</i>	Removes fitted trend from target variable.
<i>inverse_transform</i>	Adds back fitted trend to target variable.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Removes fitted trend from target variable.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the PolynomialDetrender.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns self

fit_transform (*self*, *X*, *y=None*)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the fitted trend removed.

Return type tuple of *pd.DataFrame*, *pd.Series*

inverse_transform (*self*, *y*)

Adds back fitted trend to target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable.

Returns

The first element are the input features returned without modification. The second element is the target variable y with the trend added back.

Return type tuple of *pd.DataFrame*, *pd.Series*

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Removes fitted trend from target variable.

Parameters

- **X** (*pd.DataFrame*, *optional*) – Ignored.
- **y** (*pd.Series*) – Target variable to detrend.

Returns

The input features are returned without modification. The target variable *y* is detrended

Return type tuple of *pd.DataFrame*, *pd.Series*

```
class evalml.pipelines.components.ProphetRegressor (date_index=None, change-  
                                                    point_prior_scale=0.05, sea-  
                                                    sonality_prior_scale=10, hol-  
                                                    idays_prior_scale=10, sea-  
                                                    sonality_mode='additive',  
                                                    random_seed=0,  
                                                    stan_backend='CMDSTANPY',  
                                                    **kwargs)
```

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

More information here: <https://facebook.github.io/prophet/>

Attributes

hyper-parameter_ranges	{ "changepoint_prior_scale": Real(0.001, 0.5), "seasonality_prior_scale": Real(0.01, 10), "holidays_prior_scale": Real(0.01, 10), "seasonality_mode": ["additive", "multiplicative"], }
model_family	ModelFamily.PROPHET
modifies_features	True
modifies_target	False
name	Prophet Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>build_prophet_df</i>	
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.
<i>fit</i>	Fits component to data
<i>get_params</i>	
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

static build_prophet_df (*X*, *y=None*, *date_column='ds'*)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

get_params (*self*)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X, y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.RandomForestClassifier (n_estimators=100,  
                                                    max_depth=6, n_jobs=-  
                                                    1, random_seed=0,  
                                                    **kwargs)
```

Random Forest Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 10), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Classifier
predict Uses y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (`pd.DataFrame`, or `np.ndarray`) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

```
class evalml.pipelines.components.RandomForestRegressor (n_estimators=100,
                                                         max_depth=6, n_jobs=-
                                                         1, random_seed=0,
                                                         **kwargs)
```

Random Forest Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 32), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.

continues on next page

Table 576 – continued from previous page

<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.RFClassifierSelectFromModel (number_features=None,
                                                             n_estimators=10,
                                                             max_depth=None,
                                                             per-
                                                             cent_features=0.5,
                                                             threshold=-
                                                             np.inf, n_jobs=-
                                                             1, ran-
                                                             dom_seed=0,
                                                             **kwargs)
```

Selects top features based on importance weights using a Random Forest classifier.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5.

- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to `-np.inf`.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. `-1` uses all processes. Defaults to `-1`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to `0`.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Classifier Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an <code>MethodPropertyNotFoundError</code> exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.components.RFRegressorSelectFromModel (number_features=None,  
                                                             n_estimators=10,  
                                                             max_depth=None,  
                                                             per-  
                                                             cent_features=0.5,  
                                                             threshold=- np.inf,  
                                                             n_jobs=- 1, ran-  
                                                             dom_seed=0,  
                                                             **kwargs)
```

Selects top features based on importance weights using a Random Forest regressor.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both `percent_features` and `number_features` are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to `-np.inf`.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Regressor Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type *list[str]*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an *MethodPropertyNotFoundError* exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.

- **y** (*pd.Series, optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

class `evalml.pipelines.components.SelectByType` (*column_types=None, random_seed=0, **kwargs*)

Selects columns by specified Woodwork logical type or semantic tag in input data.

Parameters

- **column_types** (*string, ww.LogicalType, list(string), list(ww.LogicalType)*) – List of Woodwork types or tags, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Select Columns By Type Transformer
needs_fitting	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the transformer by checking if column names are present in the dataset.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type `dict`

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **x** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series*, *optional*) – Targets.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type `pd.DataFrame`

class `evalml.pipelines.components.SelectColumns` (*columns=None*, *random_seed=0*,
***kwargs*)

Selects specified columns in input data.

Parameters

- **columns** (*list(string)*) – List of column names, used to determine which columns to select.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Select Columns Transformer
needs_fitting	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the transformer by checking if column names are present in the dataset.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by selecting columns.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type `dict`

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits the transformer by checking if column names are present in the dataset.

Parameters

- **X** (*pd.DataFrame*) – Data to check.
- **y** (*pd.Series, optional*) – Targets.

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by selecting columns.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Targets.

Returns Transformed X.

Return type pd.DataFrame


```
class evalml.pipelines.components.SimpleImputer (impute_strategy='most_frequent',  
                                                fill_value=None,      random_seed=0,  
                                                **kwargs)
```

Imputes missing data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types.
- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Simple Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. 'None' values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – the target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms input by imputing missing values. 'None' and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.components.SklearnStackedEnsembleClassifier (input_pipelines=None,  
                                                                    fi-  
                                                                    nal_estimator=None,  
                                                                    cv=None,  
                                                                    n_jobs=-  
                                                                    1,    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Scikit-learn Stacked Ensemble Classifier.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The classifier used to combine the base estimators. If None, uses LogisticRegressionClassifier.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, if the estimator is a classifier and y is either binary or multiclass, StratifiedKFold is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.SklearnStackedEnsembleRegressor (input_pipelines=None,  
                                                                    fi-  
                                                                    nal_estimator=None,  
                                                                    cv=None,  
                                                                    n_jobs=-  
                                                                    1,      ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

Scikit-learn Stacked Ensemble Regressor.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The regressor used to combine the base estimators. If None, uses LinearRegressor.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, KFold is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for stacked ensemble classes.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.**StandardScaler** (*random_seed=0*, ***kwargs*)

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Standard Scaler

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.components.SVMClassifier (C=1.0, kernel='rbf', gamma='auto',  
probability=True, random_seed=0,  
**kwargs)
```

Support Vector Machine Classifier.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses 1 / (n_features * X.var()) as value of gamma - If “auto” (default), uses 1 / n_features
- **probability** (*boolean*) – Whether to enable probability estimates. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0, 10), "kernel": ["poly", "rbf", "sigmoid"], "gamma": ["scale", "auto"], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Classifier
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.components.SVMRegressor (*C=1.0, kernel='rbf', gamma='auto', random_seed=0, **kwargs*)

Support Vector Machine Regressor.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses $1 / (n_features * X.var())$ as value of gamma - If “auto” (default), uses $1 / n_features$
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.TargetEncoder (cols=None, smoothing=1.0,
                                                handle_unknown='value', handle_missing='value', random_seed=0,
                                                **kwargs)
```

A transformer that encodes categorical features into target encodings.

Parameters

- **cols** (*list*) – Columns to encode. If None, all string columns will be encoded, otherwise only the columns provided will be encoded. Defaults to None
- **smoothing** (*float*) – The smoothing factor to apply. The larger this value is, the more influence the expected target value has on the resulting target encodings. Must be strictly larger than 0. Defaults to 1.0
- **handle_unknown** (*string*) – Determines how to handle unknown categories for a feature encountered. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **handle_missing** (*string*) – Determines how to handle missing values encountered during *fit* or *transform*. Options are ‘value’, ‘error’, and ‘return_nan’. Defaults to ‘value’, which replaces with the target mean
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Target Encoder

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters

continues on next page

Table 587 – continued from previous page

<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_feature_names</i>	Return feature names for the input features after fitting.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.**default_parameters** (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.**Returns** default parameters for this component.**Return type** dict**describe** (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary**Return type** None or dict**fit** (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self**fit_transform** (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

get_feature_names (*self*)

Return feature names for the input features after fitting.

Returns The feature names after encoding

Return type `np.array`

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns `ComponentBase` object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed *X*

Return type `pd.DataFrame`

```
class evalml.pipelines.components.TargetImputer (impute_strategy='most_frequent',  
                                                fill_value=None,      random_seed=0,  
                                                **kwargs)
```

Imputes missing target data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types. Defaults to “most_frequent”.
- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to `None` which uses 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "impute_strategy": ["mean", "median", "most_frequent"] }
model_family	ModelFamily.NONE
modifies_features	False
modifies_target	True
name	Target Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to target data. 'None' values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on and transforms the input target data.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input target data by imputing missing values. 'None' and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits imputer to target data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]. Ignored.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples].

Returns *self*

fit_transform (*self*, *X*, *y*)

Fits on and transforms the input target data.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.
- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (*pd.DataFrame*, *pd.Series*)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*)

Transforms input target data by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Features. Ignored.
- **y** (*pd.Series*) – Target data to impute.

Returns The original X, transformed y

Return type (*pd.DataFrame*, *pd.Series*)

class evalml.pipelines.components.**TextFeaturizer** (*random_seed=0, **kwargs*)

Transformer that can automatically featurize text columns using featuretools' nlp_primitives.

Since models cannot handle non-numeric data, any text must be broken down into features that provide useful information about that text. This component splits each text column into several informative features: Diversity Score, Mean Characters per Word, Polarity Score, and LSA (Latent Semantic Analysis). Calling transform on this component will replace any text columns in the given dataset with these numeric columns.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Text Featurization Component

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X by creating new features using existing text columns

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that Component.default_parameters == Component().parameters.

Returns default parameters for this component.

Return type dict

describe (*self, print_name=False, return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X by creating new features using existing text columns

Parameters

- **X** (*pd.DataFrame*) – The data to transform.

- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.components.TimeSeriesBaselineEstimator(gap=1, forecast_horizon=1,
random_seed=0,
**kwargs)
```

Time series estimator that predicts using the naive forecasting approach.

This is useful as a simple baseline estimator for time series problems.

Parameters

- **gap** (*int*) – Gap between prediction date and target date and must be a positive integer. If gap is 0, target date will be shifted ahead by 1 time period. Defaults to 1.
- **forecast_horizon** (*int*) – Number of time steps the model is expected to predict.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.BASELINE
modifies_features	True
modifies_target	False
name	Time Series Baseline Estimator
predict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Since baseline estimators do not use input features to calculate predictions, returns an array of zeroes.

Returns an array of zeroes

Return type np.ndarray (float)

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class `evalml.pipelines.components.Transformer` (*parameters=None*, *component_obj=None*, *random_seed=0*, ***kwargs*)

A component that may or may not need fitting that transforms data. These components are used before an estimator.

To implement a new Transformer, define your own class which is a subclass of Transformer, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Transformer component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.

continues on next page

Table 591 – continued from previous page

<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns `ComponentBase` object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y*=`None`)

Transforms data *X*.

Parameters

- **X** (`pd.DataFrame`) – Data to transform.
- **y** (`pd.Series`, *optional*) – Target data.

Returns Transformed *X*

Return type `pd.DataFrame`

```
class evalml.pipelines.components.Undersampler (sampling_ratio=0.25, sam-  
                                           pling_ratio_dict=None,  
                                           min_samples=100,  
                                           min_percentage=0.1, random_seed=0,  
                                           **kwargs)
```

Initializes an undersampling transformer to downsample the majority classes in the dataset.

This component is only run during training and not during predict.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. For instance, in a binary case where class 1 is the minority, we could specify: `sampling_ratio_dict={0: 0.5, 1: 1}`, which means we would undersample class 0 to have twice the number of samples as class 1 (minority:majority ratio = 0.5), and don’t sample class 1. Overrides `sampling_ratio` if provided. Defaults to `None`.

- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below min_percentage. Must be greater than 0. Defaults to 100.
- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above min_samples. If min_percentage and min_samples are not met, treat this as severely imbalanced, and we will not resample the data. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	True
name	Undersampler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the sampler to the data.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms the input data by sampling the data.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y*)

Fits the sampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Input features.
- **y** (*pd.Series*) – Target.

Returns self

fit_transform (*self, X, y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms the input data by sampling the data.

Parameters

- **X** (*pd.DataFrame*) – Training features.

- **y** (*pd.Series*) – Target.

Returns Transformed features and target.

Return type *pd.DataFrame*, *pd.Series*

class `evalml.pipelines.components.URLFeaturizer` (*random_seed=0*, ***kwargs*)

Transformer that can automatically extract features from URL.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	URL Featurizer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type *dict*

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component

- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.components.XGBoostClassifier(eta=0.1, max_depth=6,
                                                    min_child_weight=1,
                                                    n_estimators=100,
                                                    random_seed=0,
                                                    eval_metric='logloss',
                                                    n_jobs=12, **kwargs)
```

XGBoost Classifier.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 10), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Classifier
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path

continues on next page

Table 594 – continued from previous page

<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.components.XGBoostRegressor (eta=0.1, max_depth=6,
                                                    min_child_weight=1,
                                                    n_estimators=100, random_seed=0,
                                                    n_jobs=12,
                                                    **kwargs)
```

XGBoost Regressor.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 20), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Regressor
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

Submodules

binary_classification_pipeline

Module Contents

Classes Summary

<i>BinaryClassificationPipeline</i>	Pipeline subclass for all binary classification pipelines.
-------------------------------------	--

Contents

class evalml.pipelines.binary_classification_pipeline.**BinaryClassificationPipeline** (*component*
pa-
ram-
e-
ters=None,
cus-
tom_name=
ran-
dom_seed=

Pipeline subclass for all binary classification pipelines.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.BINARY
---------------------	---------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.

continues on next page

Table 597 – continued from previous page

<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Build a classification model. For string and categorical targets, classes are sorted
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline's feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>optimize_threshold</code>	Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels. Assumes that the column at index 1 represents the positive label case.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>threshold</code>	Threshold used to make a prediction. Defaults to None.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))` and then are mapped to values between 0 and `n_classes-1`.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [*n_samples*, *n_features*]
- **y** (*pd.Series*, *np.ndarray*) – The target training labels of length [*n_samples*]

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters `custom_hyperparameters` (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters `filepath` (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters `importance_threshold` (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are PolynomialDetrender, LabelEncoder (tbd).

Parameters `y` (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters `file_path` (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

optimize_threshold (*self*, *X*, *y*, *y_pred_proba*, *objective*)

Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.

Parameters

- **X** (*pd.DataFrame*) – Input features
- **y** (*pd.Series*) – Input target values
- **y_pred_proba** (*pd.Series*) – The predicted probabilities of the target outputted by the pipeline
- **objective** (*ObjectiveBase*) – The objective to threshold with. Must have a tunable threshold.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.**Return type** dict**predict** (*self, X, objective=None, X_train=None, y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features]
- **objective** (*Object or string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels**Return type** *pd.Series***predict_proba** (*self, X, X_train=None, y_train=None*)

Make probability estimates for labels. Assumes that the column at index 1 represents the positive label case.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates**Return type** *pd.Series***save** (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, or *np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame* or *np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

property threshold (*self*)

Threshold used to make a prediction. Defaults to None.

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

binary_classification_pipeline_mixin

Module Contents

Classes Summary

BinaryClassificationPipelineMixin

Contents

`class evalml.pipelines.binary_classification_pipeline_mixin.BinaryClassificationPipelineMixin`
Methods

<code>optimize_threshold</code>	Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.
<code>threshold</code>	Threshold used to make a prediction. Defaults to None.

`optimize_threshold(self, X, y, y_pred_proba, objective)`
Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.

Parameters

- `X` (`pd.DataFrame`) – Input features
- `y` (`pd.Series`) – Input target values
- `y_pred_proba` (`pd.Series`) – The predicted probabilities of the target outputted by the pipeline
- `objective` (`ObjectiveBase`) – The objective to threshold with. Must have a tunable threshold.

`property threshold(self)`
Threshold used to make a prediction. Defaults to None.

classification_pipeline

Module Contents

Classes Summary

<code>ClassificationPipeline</code>	Pipeline subclass for all classification pipelines.
-------------------------------------	---

Contents

`class evalml.pipelines.classification_pipeline.ClassificationPipeline(component_graph, parameters=None, custom_name=None, random_seed=0)`
Pipeline subclass for all classification pipelines.

Parameters

- `component_graph` (`list` or `dict`) – List of components in order. Accepts strings

or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component's index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names ["Imputer", "One Hot Encoder", "Imputer_2", "Logistic Regression Classifier"]

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a classification model. For string and categorical targets, classes are sorted
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component inverse_transform methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Make predictions using selected features.

continues on next page

Table 601 – continued from previous page

<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))` and then are mapped to values between 0 and `n_classes-1`.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape `[n_samples, n_features]`
- **y** (*pd.Series*, *np.ndarray*) – The target training labels of length `[n_samples]`

Returns `self`

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type `dict`

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to `None` (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **objective** (*Object* or *string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels

Return type pd.Series

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Make probability estimates for labels.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates

Return type pd.DataFrame

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, or *np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame* or *np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

component_graph

Module Contents

Classes Summary

ComponentGraph

Component graph for a pipeline as a directed acyclic graph (DAG).

Contents

class evalml.pipelines.component_graph.**ComponentGraph**(*component_dict=None*, *random_seed=0*)

Component graph for a pipeline as a directed acyclic graph (DAG).

Parameters

- **component_dict** (*dict*) – A dictionary which specifies the components and edges between components that should be used to create the component graph. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Example

```
>>> component_dict = {'imputer': ['Imputer'], 'ohe': ['One Hot Encoder', 'imputer.
↪x'], 'estimator_1': ['Random Forest Classifier', 'ohe.x'], 'estimator_2': [
↪'Decision Tree Classifier', 'ohe.x'], 'final': ['Logistic Regression Classifier
↪', 'estimator_1', 'estimator_2']}
>>> component_graph = ComponentGraph(component_dict)
```

Methods

<code>compute_final_component_features</code>	Transform all components save the final one, and gathers the data from any number of parents
<code>compute_order</code>	The order that components will be computed or called in.
<code>default_parameters</code>	The default parameter dictionary for this pipeline.
<code>describe</code>	Outputs component graph details including component parameters
<code>fit</code>	Fit each component in the graph.
<code>fit_features</code>	Fit all components save the final one, usually an estimator.
<code>generate_order</code>	Regenerated the topologically sorted order of the graph
<code>get_component</code>	Retrieves a single component object from the graph.
<code>get_estimators</code>	Gets a list of all the estimator components within this graph.
<code>get_inputs</code>	Retrieves all inputs for a given component.
<code>get_last_component</code>	Retrieves the component that is computed last in the graph, usually the final estimator.
<code>graph</code>	Generate an image representing the component graph
<code>instantiate</code>	Instantiates all uninstantiated components within the graph using the given parameters. An error will be
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>predict</code>	Make predictions using selected features.
<code>transform</code>	Transform the input using the component graph.

compute_final_component_features (*self*, *X*, *y=None*)

Transform all components save the final one, and gathers the data from any number of parents to get all the information that should be fed to the final component.

Parameters

- **x** (*pd.DataFrame*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples]. Defaults to None.

Returns Transformed values.

Return type *pd.DataFrame*

property compute_order (*self*)

The order that components will be computed or called in.

property default_parameters (*self*)

The default parameter dictionary for this pipeline.

Returns Dictionary of all component default parameters.

Return type *dict*

describe (*self*, *return_dict=False*)

Outputs component graph details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about component graph. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

fit (*self*, *X*, *y*)

Fit each component in the graph.

Parameters

- **x** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].

fit_features (*self*, *X*, *y*)

Fit all components save the final one, usually an estimator.

Parameters

- **x** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].

Returns Transformed values.

Return type *pd.DataFrame*

classmethod generate_order (*cls*, *component_dict*)

Regenerated the topologically sorted order of the graph

get_component (*self*, *component_name*)

Retrieves a single component object from the graph.

Parameters **component_name** (*str*) – Name of the component to retrieve

Returns *ComponentBase* object

get_estimators (*self*)

Gets a list of all the estimator components within this graph.

Returns All estimator objects within the graph.

Return type *list*

get_inputs (*self*, *component_name*)

Retrieves all inputs for a given component.

Parameters **component_name** (*str*) – Name of the component to look up.

Returns List of inputs for the component to use.

Return type list[str]

get_last_component (*self*)

Retrieves the component that is computed last in the graph, usually the final estimator.

Returns ComponentBase object

graph (*self*, *name=None*, *graph_format=None*)

Generate an image representing the component graph

Parameters

- **name** (*str*) – Name of the graph. Defaults to None.
- **graph_format** (*str*) – file format to save the graph in. Defaults to None.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

instantiate (*self*, *parameters*)

Instantiates all uninstantiated components within the graph using the given parameters. An error will be raised if a component is already instantiated but the parameters dict contains arguments for that component.

Parameters **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary {} or None implies using all default values for component parameters. If a component in the component graph is already instantiated, it will not use any of its parameters defined in this dictionary.

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** – (pd.Series): Final component features

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*) – Input features of shape [n_samples, n_features].

Returns Predicted values.

Return type pd.Series

transform (*self*, *X*, *y=None*)

Transform the input using the component graph.

Parameters

- **X** (*pd.DataFrame*) – Input features of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

multiclass_classification_pipeline

Module Contents

Classes Summary

<i>MulticlassClassificationPipeline</i>	Pipeline subclass for all multiclass classification pipelines.
---	--

Contents

class evalml.pipelines.multiclass_classification_pipeline.**MulticlassClassificationPipeline**

Pipeline subclass for all multiclass classification pipelines.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.MULTICLASS
---------------------	-------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.

continues on next page

Table 605 – continued from previous page

<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Build a classification model. For string and categorical targets, classes are sorted
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline's feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))` and then are mapped to values between 0 and `n_classes-1`.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape `[n_samples, n_features]`
- **y** (*pd.Series*, *np.ndarray*) – The target training labels of length `[n_samples]`

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type *Component*

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type *dict*

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to `None` (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (`tbd`).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns `PipelineBase` object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or `None` implies using all default values for component parameters. Defaults to `None`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type `dict`

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features]
- **objective** (*Object or string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self, X, X_train=None, y_train=None*)

Make probability estimates for labels.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates

Return type *pd.DataFrame*

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self, X, y, objectives, X_train=None, y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series, or np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame or np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

pipeline_base**Module Contents****Classes Summary**

PipelineBase

Machine learning pipeline made out of transformers and an Estimator.

Attributes Summary

logger

Contents

evalml.pipelines.pipeline_base.**logger**

class evalml.pipelines.pipeline_base.**PipelineBase** (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Machine learning pipeline made out of transformers and an Estimator.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”].
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a model.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Make predictions using selected features.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on current and additional objectives.
<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

abstract fit (*self*, *X*, *y*)

Build a model.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training data of length [n_samples].

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Predicted values.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

abstract score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores.

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

pipeline_meta

Module Contents

Classes Summary

<i>PipelineBaseMeta</i>	Metaclass that overrides creating a new pipeline by wrapping methods with validators and setters
-------------------------	--

Contents

class evalml.pipelines.pipeline_meta.**PipelineBaseMeta**

Metaclass that overrides creating a new pipeline by wrapping methods with validators and setters

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	['predict', 'predict_proba', 'transform', 'inverse_transform']
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<i>check_for_fit</i>	<i>check_for_fit</i> wraps a method that validates if <i>self._is_fitted</i> is <i>True</i> .
<i>register</i>	Register a virtual subclass of an ABC.
<i>set_fit</i>	

classmethod **check_for_fit** (*cls*, *method*)

check_for_fit wraps a method that validates if *self._is_fitted* is *True*. It raises an exception if *False* and calls and returns the wrapped method if *True*.

register (*cls*, *subclass*)

Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

classmethod **set_fit** (*cls*, *method*)

regression_pipeline

Module Contents

Classes Summary

<i>RegressionPipeline</i>	Pipeline subclass for all regression pipelines.
---------------------------	---

Contents

class evalml.pipelines.regression_pipeline.**RegressionPipeline** (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Pipeline subclass for all regression pipelines.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.REGRESSION
---------------------	-------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a regression model.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance
<i>inverse_transform</i>	Apply component inverse_transform methods to estimator predictions in reverse order.

continues on next page

Table 612 – continued from previous page

<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property `feature_importance` (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns `pd.DataFrame` including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a regression model.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – The input training data of shape `[n_samples, n_features]`
- **y** (`pd.Series`, `np.ndarray`) – The target training data of length `[n_samples]`

Returns `self`

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type `dict`

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to `None` (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (tbd).

Parameters **y** (`pd.Series`) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)
Returns model family of this pipeline.

property name (*self*)
Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)
Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)
Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Predicted values.

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)
Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)
Evaluate model performance on current and additional objectives

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]

- **y** (*pd.Series*, or *np.ndarray*) – True values of length [n_samples]
- **objectives** (*list*) – Non-empty list of objectives to score on
- **x_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **x** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

time_series_classification_pipelines

Module Contents

Classes Summary

<i>TimeSeriesBinaryClassificationPipeline</i>	Pipeline base class for time series binary classification problems.
<i>TimeSeriesClassificationPipeline</i>	Pipeline base class for time series classification problems.
<i>TimeSeriesMulticlassClassificationPipeline</i>	Pipeline base class for time series multiclass classification problems.

Contents

class evalml.pipelines.time_series_classification_pipelines.**TimeSeriesBinaryClassification**

Pipeline base class for time series binary classification problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={“pipeline”: {“date_index”: “Date”, “max_delay”: 4, “gap”: 2}}).
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance
<i>inverse_transform</i>	Apply component inverse_transform methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

continues on next page

Table 614 – continued from previous page

<code>optimize_threshold</code>	Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Predict on future data where target is not known.
<code>predict_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>predict_proba</code>	Predict on future data where the target is unknown.
<code>predict_proba_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>threshold</code>	Threshold used to make a prediction. Defaults to None.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **X_train** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type dict

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns pd.DataFrame including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series classification pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns self

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than *importance_threshold*. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters *y* (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters *file_path* (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

optimize_threshold (*self*, *X*, *y*, *y_pred_proba*, *objective*)

Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.

Parameters

- **X** (*pd.DataFrame*) – Input features
- **y** (*pd.Series*) – Input target values
- **y_pred_proba** (*pd.Series*) – The predicted probabilities of the target outputted by the pipeline
- **objective** (*ObjectiveBase*) – The objective to threshold with. Must have a tunable threshold.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].

- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self, X, y, X_train, y_train, objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase, str, None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type pd.Series

predict_proba (*self, X, X_train=None, y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type pd.Series

predict_proba_in_sample (*self, X_holdout, y_holdout, X_train, y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].

- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

property threshold (*self*)

Threshold used to make a prediction. Defaults to None.

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*


```
class evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline
```

Pipeline base class for time series classification problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as `date_index`, `gap`, and `max_delay` must be specified with the “pipeline” key. For example: `Pipeline(parameters={"pipeline": {"date_index": "Date", "max_delay": 4, "gap": 2}})`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<code>can_tune_threshold_with_objective</code>	Determine whether the threshold of a binary classification pipeline can be tuned.
<code>classes_</code>	Gets the class names for the problem.
<code>clone</code>	Constructs a new pipeline with the same components, parameters, and random state.
<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Fit a time series classification pipeline.
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.

continues on next page

Table 615 – continued from previous page

<code>graph_feature_importance</code>	Generate a bar graph of the pipeline’s feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Predict on future data where target is not known.
<code>predict_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>predict_proba</code>	Predict on future data where the target is unknown.
<code>predict_proba_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

`can_tune_threshold_with_objective` (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **`pipeline`** (*PipelineBase*) – Binary classification pipeline.
- **`objective`** – Primary AutoMLSearch objective.

`property classes_` (*self*)

Gets the class names for the problem.

`clone` (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

`compute_estimator_features` (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **`X`** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **`y`** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **`X_train`** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **`y_train`** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type `pd.DataFrame`

static create_objectives (*objectives*)

property custom_name (*self*)
Custom name of the pipeline.

describe (*self*, *return_dict=False*)
Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline.
Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type `dict`

property feature_importance (*self*)
Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns `pd.DataFrame` including feature names and their corresponding importance

fit (*self*, *X*, *y*)
Fit a time series classification pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape `[n_samples, n_features]`.
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length `[n_samples]`.

Returns *self*

get_component (*self*, *name*)
Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type `Component`

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)
Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type `dict`

graph (*self*, *filepath=None*)
Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)
Generate a bar graph of the pipeline's feature importance

Parameters `importance_threshold` (*float, optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self, y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (tbd).

Parameters `y` (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters `file_path` (*str*) – location to load file

Returns `PipelineBase` object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self, parameters, random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or `None` implies using all default values for component parameters. Defaults to `None`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type `dict`

predict (*self, X, objective=None, X_train=None, y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape `[n_samples, n_features]`.
- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self*, *X*, *y*, *X_train*, *y_train*, *objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase*, *str*, *None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type *pd.Series*

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

predict_proba_in_sample (*self*, *X_holdout*, *y_holdout*, *X_train*, *y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file

- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self, X, y, objectives, X_train=None, y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

class evalml.pipelines.time_series_classification_pipelines.**TimeSeriesMulticlassClassifier**

Pipeline base class for time series multiclass classification problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as

date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={"pipeline": {"date_index": "Date", "max_delay": 4, "gap": 2}}).

- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.TIME_SERIES_MULTICLASS
---------------------	-------------------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance
<i>inverse_transform</i>	Apply component inverse_transform methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Predict on future data where target is not known.
<i>predict_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>predict_proba</i>	Predict on future data where the target is unknown.
<i>predict_proba_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on current and additional objectives.

continues on next page

Table 616 – continued from previous page

<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **X_train** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series classification pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns self

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self, X, objective=None, X_train=None, y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self, X, y, X_train, y_train, objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase, str, None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type pd.Series

predict_proba (*self, X, X_train=None, y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

predict_proba_in_sample (*self*, *X_holdout*, *y_holdout*, *X_train*, *y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

time_series_pipeline_base**Module Contents****Classes Summary**

TimeSeriesPipelineBase

Pipeline base class for time series problems.

Contents

```
class evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase (component_graph,  
                                                                    pa-  
                                                                    ram-  
                                                                    e-  
                                                                    ters=None,  
                                                                    cus-  
                                                                    tom_name=None,  
                                                                    ran-  
                                                                    dom_seed=0)
```

Pipeline base class for time series problems.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={“pipeline”: {“date_index”: “Date”, “max_delay”: 4, “gap”: 2}}).
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<code>can_tune_threshold_with_objective</code>	Determine whether the threshold of a binary classification pipeline can be tuned.
<code>clone</code>	Constructs a new pipeline with the same components, parameters, and random state.
<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Fit a time series pipeline.
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline's feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Predict on future data where target is not known.
<code>predict_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **X_train** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self*, *X*, *y*, *X_train*, *y_train*, *objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features]
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples]
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features]
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train]
- **objective** (*ObjectiveBase*, *str*, *None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

abstract score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – True labels of length [n_samples].

- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

time_series_regression_pipeline

Module Contents

Classes Summary

TimeSeriesRegressionPipeline

Pipeline base class for time series regression problems.

Contents

class evalml.pipelines.time_series_regression_pipeline.**TimeSeriesRegressionPipeline** (*component*

*pa-
ram-
e-
ters=None
cus-
tom_name
ran-
dom_seed*

Pipeline base class for time series regression problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”,

“Logistic Regression Classifier”]

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as `date_index`, `gap`, and `max_delay` must be specified with the “pipeline” key. For example: `Pipeline(parameters={"pipeline": {"date_index": "Date", "max_delay": 4, "gap": 2}})`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.TIME_SERIES_REGRESSION
---------------------	-------------------------------------

Methods

<code>can_tune_threshold_with_objective</code>	Determine whether the threshold of a binary classification pipeline can be tuned.
<code>clone</code>	Constructs a new pipeline with the same components, parameters, and random state.
<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Fit a time series pipeline.
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline’s feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Predict on future data where target is not known.
<code>predict_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.

continues on next page

Table 620 – continued from previous page

<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **X_train** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].

- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns self

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self, X, objective=None, X_train=None, y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self, X, y, X_train, y_train, objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features]
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples]
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features]
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train]
- **objective** (*ObjectiveBase, str, None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

utils

Module Contents

Functions

<code>generate_pipeline_code</code>	Creates and returns a string that contains the Python imports and code required for running the EvalML pipeline.
<code>make_pipeline</code>	Given input data, target data, an estimator class and the problem type,
<code>make_timeseries_baseline_pipeline</code>	Make a baseline pipeline for time series regression problems.

Attributes Summary

logger

Contents

`evalml.pipelines.utils.generate_pipeline_code(element)`

Creates and returns a string that contains the Python imports and code required for running the EvalML pipeline.

Parameters `element` (*pipeline instance*) – The instance of the pipeline to generate string Python code

Returns String representation of Python code that can be run separately in order to recreate the pipeline instance. Does not include code for custom component implementation.

`evalml.pipelines.utils.logger`

`evalml.pipelines.utils.make_pipeline(X, y, estimator, problem_type, parameters=None, sampler_name=None, extra_components=None)`

Given input data, target data, an estimator class and the problem type, generates a pipeline class with a preprocessing chain which was recommended based on the inputs. The pipeline will be a subclass of the appropriate pipeline base class for the specified problem_type.

Parameters

- **X** (*pd.DataFrame*) – The input data of shape [n_samples, n_features]
- **y** (*pd.Series*) – The target data of length [n_samples]
- **estimator** (*Estimator*) – Estimator for pipeline
- **problem_type** (*ProblemTypes* or *str*) – Problem type for pipeline to generate
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters.
- **sampler_name** (*str*) – The name of the sampler component to add to the pipeline. Only used in classification problems. Defaults to None
- **extra_components** – List of extra components to be added after preprocessing components. Defaults to None.

`evalml.pipelines.utils.make_timeseries_baseline_pipeline(problem_type, gap, forecast_horizon)`

Make a baseline pipeline for time series regression problems.

Parameters

- **problem_type** – One of TIME_SERIES_REGRESSION, TIME_SERIES_MULTICLASS, TIME_SERIES_BINARY
- **gap** (*int*) – Non-negative gap parameter.
- **forecast_horizon** (*int*) – Positive forecast_horizon parameter.

Returns TimeSeriesPipelineBase, a time series pipeline corresponding to the problem type.

Package Contents

Classes Summary

<i>ARIMAREgressor</i>	Autoregressive Integrated Moving Average Model.
<i>BinaryClassificationPipeline</i>	Pipeline subclass for all binary classification pipelines.
<i>CatBoostClassifier</i>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.
<i>CatBoostRegressor</i>	CatBoost Regressor, a regressor that uses gradient-boosting on decision trees.
<i>ClassificationPipeline</i>	Pipeline subclass for all classification pipelines.
<i>ComponentGraph</i>	Component graph for a pipeline as a directed acyclic graph (DAG).
<i>DecisionTreeClassifier</i>	Decision Tree Classifier.
<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>DelayedFeatureTransformer</i>	Transformer that delays input features and target variable for time series problems.
<i>DFSTransformer</i>	Featuretools DFS component that generates features for the input features.
<i>ElasticNetClassifier</i>	Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.
<i>ElasticNetRegressor</i>	Elastic Net Regressor.
<i>Estimator</i>	A component that fits and predicts given data.
<i>ExtraTreesClassifier</i>	Extra Trees Classifier.
<i>ExtraTreesRegressor</i>	Extra Trees Regressor.
<i>FeatureSelector</i>	Selects top features based on importance weights.
<i>KNeighborsClassifier</i>	K-Nearest Neighbors Classifier.
<i>LightGBMClassifier</i>	LightGBM Classifier.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>LinearRegressor</i>	Linear Regressor.
<i>LogisticRegressionClassifier</i>	Logistic Regression Classifier.
<i>MulticlassClassificationPipeline</i>	Pipeline subclass for all multiclass classification pipelines.
<i>OneHotEncoder</i>	A transformer that encodes categorical features in a one-hot numeric array.
<i>PerColumnImputer</i>	Imputes missing data according to a specified imputation strategy per column.
<i>PipelineBase</i>	Machine learning pipeline made out of transformers and an Estimator.
<i>ProphetRegressor</i>	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
<i>RandomForestClassifier</i>	Random Forest Classifier.
<i>RandomForestRegressor</i>	Random Forest Regressor.
<i>RegressionPipeline</i>	Pipeline subclass for all regression pipelines.
<i>RFClassifierSelectFromModel</i>	Selects top features based on importance weights using a Random Forest classifier.
<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.

continues on next page

Table 623 – continued from previous page

<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>SklearnStackedEnsembleClassifier</i>	Scikit-learn Stacked Ensemble Classifier.
<i>SklearnStackedEnsembleRegressor</i>	Scikit-learn Stacked Ensemble Regressor.
<i>StandardScaler</i>	A transformer that standardizes input features by removing the mean and scaling to unit variance.
<i>SVMClassifier</i>	Support Vector Machine Classifier.
<i>SVMRegressor</i>	Support Vector Machine Regressor.
<i>TargetEncoder</i>	A transformer that encodes categorical features into target encodings.
<i>TimeSeriesBinaryClassificationPipeline</i>	Pipeline base class for time series binary classification problems.
<i>TimeSeriesClassificationPipeline</i>	Pipeline base class for time series classification problems.
<i>TimeSeriesMulticlassClassificationPipeline</i>	Pipeline base class for time series multiclass classification problems.
<i>TimeSeriesRegressionPipeline</i>	Pipeline base class for time series regression problems.
<i>Transformer</i>	A component that may or may not need fitting that transforms data.
<i>XGBoostClassifier</i>	XGBoost Classifier.
<i>XGBoostRegressor</i>	XGBoost Regressor.

Contents

class evalml.pipelines.**ARIMAREgressor** (*date_index=None, trend=None, start_p=2, d=0, start_q=2, max_p=5, max_d=2, max_q=5, seasonal=True, n_jobs=-1, random_seed=0, **kwargs*)

Autoregressive Integrated Moving Average Model. The three parameters (p, d, q) are the AR order, the degree of differencing, and the MA order. More information here: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

Currently ARIMAREgressor isn't supported via conda install. It's recommended that it be installed via PyPI.

Parameters

- **date_index** (*str*) – Specifies the name of the column in X that provides the datetime objects. Defaults to None.
- **trend** (*str*) – Controls the deterministic trend. Options are ['n', 'c', 't', 'ct'] where 'c' is a constant term, 't' indicates a linear trend, and 'ct' is both. Can also be an iterable when defining a polynomial, such as [1, 1, 0, 1].
- **start_p** (*int*) – Minimum Autoregressive order. Defaults to 2.
- **d** (*int*) – Minimum Differencing degree. Defaults to 0.
- **start_q** (*int*) – Minimum Moving Average order. Defaults to 2.
- **max_p** (*int*) – Maximum Autoregressive order. Defaults to 5.
- **max_d** (*int*) – Maximum Differencing degree. Defaults to 2.
- **max_q** (*int*) – Maximum Moving Average order. Defaults to 5.
- **seasonal** (*boolean*) – Whether to fit a seasonal model to ARIMA. Defaults to True.

- **n_jobs** (*int* or *None*) – Non-negative integer describing level of parallelism used for pipelines. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "start_p": Integer(1, 3), "d": Integer(0, 2), "start_q": Integer(1, 3), "max_p": Integer(3, 10), "max_d": Integer(2, 5), "max_q": Integer(3, 10), "seasonal": [True, False], }
model_family	ModelFamily.ARIMA
modifies_features	True
modifies_target	False
name	ARIMA Regressor
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X, y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.BinaryClassificationPipeline(component_graph, parameters=None, custom_name=None, random_seed=0)
```

Pipeline subclass for all binary classification pipelines.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.BINARY
---------------------	---------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a classification model. For string and categorical targets, classes are sorted
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance
<i>inverse_transform</i>	Apply component inverse_transform methods to estimator predictions in reverse order.

continues on next page

Table 625 – continued from previous page

<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>optimize_threshold</code>	Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels. Assumes that the column at index 1 represents the positive label case.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>threshold</code>	Threshold used to make a prediction. Defaults to None.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (`PipelineBase`) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (`pd.DataFrame`) – Input data to the pipeline to transform.
- **y** (`pd.Series` or `None`) – Targets corresponding to X. optional.
- **X_train** (`pd.DataFrame` or `np.ndarray` or `None`) – Training data. Only used for time series.
- **y_train** (`pd.Series` or `None`) – Training labels. Only used for time series.

Returns New transformed features.

Return type `pd.DataFrame`

static create_objectives (*objectives*)

property `custom_name` (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters `return_dict` (*bool*) – If True, return dictionary of information about pipeline.
Defaults to False.

Returns Dictionary of all component parameters if `return_dict` is True, else None

Return type dict

property `feature_importance` (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns `pd.DataFrame` including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))`
and then are mapped to values between 0 and `n_classes-1`.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – The input training data of shape `[n_samples, n_features]`
- **y** (`pd.Series`, `np.ndarray`) – The target training labels of length `[n_samples]`

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters `name` (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters `custom_hyperparameters` (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters `filepath` (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters `importance_threshold` (*float, optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self, y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (tbd).

Parameters `y` (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters `file_path` (*str*) – location to load file

Returns `PipelineBase` object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self, parameters, random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or `None` implies using all default values for component parameters. Defaults to `None`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

optimize_threshold (*self, X, y, y_pred_proba, objective*)

Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.

Parameters

- **X** (*pd.DataFrame*) – Input features
- **y** (*pd.Series*) – Input target values
- **y_pred_proba** (*pd.Series*) – The predicted probabilities of the target outputted by the pipeline
- **objective** (*ObjectiveBase*) – The objective to threshold with. Must have a tunable threshold.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type `dict`

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **objective** (*Object* or *string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Make probability estimates for labels. Assumes that the column at index 1 represents the positive label case.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, or *np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame* or *np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

property threshold (*self*)

Threshold used to make a prediction. Defaults to None.

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **x** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

class evalml.pipelines.CatBoostClassifier (*n_estimators=10*, *eta=0.03*, *max_depth=6*, *bootstrap_type=None*, *silent=True*, *allow_writing_files=False*, *random_seed=0*, *n_jobs=-1*, ***kwargs*)

CatBoost Classifier, a classifier that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(4, 100), "eta": Real(0.000001, 1), "max_depth": Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.CatBoostRegressor (n_estimators=10, eta=0.03, max_depth=6,  
                                           bootstrap_type=None, silent=False, al-  
                                           low_writing_files=False, random_seed=0,  
                                           n_jobs=-1, **kwargs)
```

CatBoost Regressor, a regressor that uses gradient-boosting on decision trees. CatBoost is an open-source library and natively supports categorical features.

For more information, check out <https://catboost.ai/>

Parameters

- **n_estimators** (*float*) – The maximum number of trees to build. Defaults to 10.
- **eta** (*float*) – The learning rate. Defaults to 0.03.
- **max_depth** (*int*) – The maximum tree depth for base learners. Defaults to 6.
- **bootstrap_type** (*string*) – Defines the method for sampling the weights of objects. Available methods are ‘Bayesian’, ‘Bernoulli’, ‘MVS’. Defaults to None.
- **silent** (*boolean*) – Whether to use the “silent” logging mode. Defaults to True.
- **allow_writing_files** (*boolean*) – Whether to allow writing snapshot files while training. Defaults to False.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(4, 100), “eta”: Real(0.000001, 1), “max_depth”: Integer(4, 10), }
model_family	ModelFamily.CATBOOST
modifies_features	True
modifies_target	False
name	CatBoost Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 627 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property `parameters` (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

class `evalml.pipelines.ClassificationPipeline` (*component_graph*, *parameters=None*,
custom_name=None, *random_seed=0*)

Pipeline subclass for all classification pipelines.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or `ComponentBase` subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or `None` implies using all default values for component parameters. Defaults to `None`.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to `None`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	<code>None</code>
---------------------	-------------------

Methods

<code>can_tune_threshold_with_objective</code>	Determine whether the threshold of a binary classification pipeline can be tuned.
<code>classes_</code>	Gets the class names for the problem.
<code>clone</code>	Constructs a new pipeline with the same components, parameters, and random state.
<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Build a classification model. For string and categorical targets, classes are sorted
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline's feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

`can_tune_threshold_with_objective` (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **`pipeline`** (`PipelineBase`) – Binary classification pipeline.
- **`objective`** – Primary `AutoMLSearch` objective.

`property classes_` (*self*)

Gets the class names for the problem.

`clone` (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random

state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline.
Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))` and then are mapped to values between 0 and *n_classes*-1.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [*n_samples*, *n_features*]
- **y** (*pd.Series*, *np.ndarray*) – The target training labels of length [*n_samples*]

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters `custom_hyperparameters` (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters `filepath` (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters `importance_threshold` (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are PolynomialDetrender, LabelEncoder (tbd).

Parameters `y` (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters `file_path` (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **objective** (*Object* or *string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Make probability estimates for labels.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates

Return type *pd.DataFrame*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, or *np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame* or *np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [*n_samples*, *n_features*].
- **y** (*pd.Series*) – The target data of length [*n_samples*]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

class evalml.pipelines.**ComponentGraph** (*component_dict=None*, *random_seed=0*)

Component graph for a pipeline as a directed acyclic graph (DAG).

Parameters

- **component_dict** (*dict*) – A dictionary which specifies the components and edges between components that should be used to create the component graph. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Example

```
>>> component_dict = {'imputer': ['Imputer'], 'ohe': ['One Hot Encoder', 'imputer.
↪x'], 'estimator_1': ['Random Forest Classifier', 'ohe.x'], 'estimator_2': [
↪'Decision Tree Classifier', 'ohe.x'], 'final': ['Logistic Regression Classifier
↪', 'estimator_1', 'estimator_2']}
>>> component_graph = ComponentGraph(component_dict)
```

Methods

<i>compute_final_component_features</i>	Transform all components save the final one, and gathers the data from any number of parents
<i>compute_order</i>	The order that components will be computed or called in.
<i>default_parameters</i>	The default parameter dictionary for this pipeline.
<i>describe</i>	Outputs component graph details including component parameters
<i>fit</i>	Fit each component in the graph.
<i>fit_features</i>	Fit all components save the final one, usually an estimator.
<i>generate_order</i>	Regenerated the topologically sorted order of the graph
<i>get_component</i>	Retrieves a single component object from the graph.
<i>get_estimators</i>	Gets a list of all the estimator components within this graph.
<i>get_inputs</i>	Retrieves all inputs for a given component.
<i>get_last_component</i>	Retrieves the component that is computed last in the graph, usually the final estimator.
<i>graph</i>	Generate an image representing the component graph

continues on next page

Table 629 – continued from previous page

<i>instantiate</i>	Instantiates all uninstantiated components within the graph using the given parameters. An error will be
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>predict</i>	Make predictions using selected features.
<i>transform</i>	Transform the input using the component graph.

compute_final_component_features (*self*, *X*, *y=None*)

Transform all components save the final one, and gathers the data from any number of parents to get all the information that should be fed to the final component.

Parameters

- **X** (*pd.DataFrame*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples]. Defaults to None.

Returns Transformed values.

Return type *pd.DataFrame*

property compute_order (*self*)

The order that components will be computed or called in.

property default_parameters (*self*)

The default parameter dictionary for this pipeline.

Returns Dictionary of all component default parameters.

Return type *dict*

describe (*self*, *return_dict=False*)

Outputs component graph details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about component graph. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

fit (*self*, *X*, *y*)

Fit each component in the graph.

Parameters

- **X** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].

fit_features (*self*, *X*, *y*)

Fit all components save the final one, usually an estimator.

Parameters

- **X** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target training data of length [n_samples].

Returns Transformed values.

Return type *pd.DataFrame*

classmethod generate_order (*cls*, *component_dict*)

Regenerated the topologically sorted order of the graph

get_component (*self*, *component_name*)

Retrieves a single component object from the graph.

Parameters *component_name* (*str*) – Name of the component to retrieve

Returns ComponentBase object

get_estimators (*self*)

Gets a list of all the estimator components within this graph.

Returns All estimator objects within the graph.

Return type list

get_inputs (*self*, *component_name*)

Retrieves all inputs for a given component.

Parameters *component_name* (*str*) – Name of the component to look up.

Returns List of inputs for the component to use.

Return type list[str]

get_last_component (*self*)

Retrieves the component that is computed last in the graph, usually the final estimator.

Returns ComponentBase object

graph (*self*, *name=None*, *graph_format=None*)

Generate an image representing the component graph

Parameters

- **name** (*str*) – Name of the graph. Defaults to None.
- **graph_format** (*str*) – file format to save the graph in. Defaults to None.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

instantiate (*self*, *parameters*)

Instantiates all uninstantiated components within the graph using the given parameters. An error will be raised if a component is already instantiated but the parameters dict contains arguments for that component.

Parameters *parameters* (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary {} or None implies using all default values for component parameters. If a component in the component graph is already instantiated, it will not use any of its parameters defined in this dictionary.

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters *y* – (pd.Series): Final component features

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*) – Input features of shape [n_samples, n_features].

Returns Predicted values.

Return type pd.Series

transform (*self*, *X*, *y=None*)

Transform the input using the component graph.

Parameters

- **X** (*pd.DataFrame*) – Input features of shape [*n_samples*, *n_features*].
- **y** (*pd.Series*) – The target data of length [*n_samples*]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

```
class evalml.pipelines.DecisionTreeClassifier (criterion='gini',      max_features='auto',  
                                              max_depth=6,      min_samples_split=2,  
                                              min_weight_fraction_leaf=0.0,      ran-  
                                              dom_seed=0, **kwargs)
```

Decision Tree Classifier.

Parameters

- **criterion** (*{ "gini", "entropy" }*) – The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Defaults to “gini”.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and *int(max_features * n_features)* features are considered at each split.
 - If “auto”, then *max_features=sqrt(n_features)*.
 - If “sqrt”, then *max_features=sqrt(n_features)*.
 - If “log2”, then *max_features=log2(n_features)*.
 - If *None*, then *max_features = n_features*.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and *ceil(min_samples_split * n_samples)* are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["gini", "entropy"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None


```
class evalml.pipelines.DecisionTreeRegressor (criterion='mse', max_features='auto',  

max_depth=6, min_samples_split=2,  

min_weight_fraction_leaf=0.0, random_seed=0, **kwargs)
```

Decision Tree Regressor.

Parameters

- **criterion** (*{ "mse", "friedman_mse", "mae", "poisson" }*) – The function to measure the quality of a split. Supported criteria are:
 - “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node
 - “friedman_mse”, which uses mean squared error with Friedman’s improvement score for potential splits
 - “mae” for the mean absolute error, which minimizes the L1 loss using the median of each terminal node,
 - “poisson” which uses reduction in Poisson deviance to find splits.
- **max_features** (*int, float or { "auto", "sqrt", "log2" }*) – The number of features to consider when looking for the best split:
 - If int, then consider max_features features at each split.
 - If float, then max_features is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “sqrt”, then $\text{max_features} = \sqrt{\text{n_features}}$.
 - If “log2”, then $\text{max_features} = \log_2(\text{n_features})$.
 - If None, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If int, then consider min_samples_split as the minimum number.
 - If float, then min_samples_split is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.

Defaults to 2.

- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "criterion": ["mse", "friedman_mse", "mae"], "max_features": ["auto", "sqrt", "log2"], "max_depth": Integer(4, 10), }
model_family	ModelFamily.DECISION_TREE
modifies_features	True
modifies_target	False
name	Decision Tree Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.DelayedFeatureTransformer(date_index=None, max_delay=2,  
                                                gap=0, forecast_horizon=1,  
                                                delay_features=True, de-  
                                                lay_target=True, random_seed=0,  
                                                **kwargs)
```

Transformer that delays input features and target variable for time series problems.

Parameters

- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Ignored.
- **max_delay** (*int*) – Maximum number of time units to delay each feature. Defaults to 2.
- **forecast_horizon** (*int*) – The number of time periods the pipeline is expected to forecast.
- **delay_features** (*bool*) – Whether to delay the input features. Defaults to True.
- **delay_target** (*bool*) – Whether to delay the target. Defaults to True.
- **gap** (*int*) – The number of time units between when the features are collected and when the target is collected. For example, if you are predicting the next time step's target, gap=1. This is only needed because when gap=0, we need to be sure to start the lagging of the target variable at 1. Defaults to 1.
- **random_seed** (*int*) – Seed for the random number generator. This transformer performs the same regardless of the random seed provided.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Delayed Feature Transformer
needs_fitting	False
target_get_colname_prefix	target_delay_{}

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits the DelayFeatureTransformer.
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path

continues on next page

Table 632 – continued from previous page

<i>transform</i>	Computes the delayed features for all features in X and y.
------------------	--

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DelayFeatureTransformer.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Computes the delayed features for all features in X and y.

For each feature in X, it will add a column to the output dataframe for each delay in the (inclusive) range [1, max_delay]. The values of each delayed feature are simply the original feature shifted forward in time by the delay amount. For example, a delay of 3 units means that the feature value at row n will be taken from the n-3rd row of that feature

If y is not None, it will also compute the delayed values for the target variable.

Parameters

- **X** (*pd.DataFrame* or *None*) – Data to transform. None is expected when only the target variable is being used.
- **y** (*pd.Series*, or *None*) – Target.

Returns Transformed X.

Return type *pd.DataFrame*

class evalml.pipelines.DFSTransformer (*index='index'*, *random_seed=0*, ***kwargs*)

Featuretools DFS component that generates features for the input features.

Parameters

- **index** (*string*) – The name of the column that contains the indices. If no column with this name exists, then featuretools.EntitySet() creates a column with this name to serve as the index column. Defaults to 'index'.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS Transformer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
--------------	---

<i>default_parameters</i>	Returns the default parameters for this component.
---------------------------	--

continues on next page

Table 633 – continued from previous page

<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits the DFSTransformer Transformer component.
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Computes the feature matrix for the input X using featuretools’ dfs algorithm.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits the DFSTransformer Transformer component.

Parameters

- **X** (*pd.DataFrame*, *np.array*) – The input data to transform, of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type `pd.DataFrame`

static load (*file_path*)

Loads component at file path

Parameters `file_path` (*str*) – Location to load file

Returns `ComponentBase` object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- `file_path` (*str*) – Location to save file
- `pickle_protocol` (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y=None*)

Computes the feature matrix for the input *X* using featuretools' dfs algorithm.

Parameters

- `X` (*pd.DataFrame* or *np.ndarray*) – The input training data to transform. Has shape `[n_samples, n_features]`
- `y` (*pd.Series*, *optional*) – Ignored.

Returns Feature matrix

Return type `pd.DataFrame`

class `evalml.pipelines.ElasticNetClassifier` (*penalty='elasticnet'*, *C=1.0*, *l1_ratio=0.15*, *multi_class='auto'*, *solver='saga'*, *n_jobs=-1*, *random_seed=0*, ***kwargs*)

Elastic Net Classifier. Uses Logistic Regression with elasticnet penalty as the base estimator.

Parameters

- **penalty** (`{"l1", "l2", "elasticnet", "none"}`) – The norm used in penalization. Defaults to “elasticnet”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **multi_class** (`{"auto", "ovr", "multinomial"}`) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when `solver='liblinear'`. “auto” selects “ovr” if the data is binary, or if `solver='liblinear'`, and otherwise selects “multinomial”. Defaults to “auto”.

- **solver** (`{"newton-cg", "lbfgs", "liblinear", "sag", "saga"}`) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only “newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.
 - “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
 - “liblinear” and “saga” also handle L1 penalty
 - “saga” also supports “elasticnet” penalty
 - “liblinear” does not support setting `penalty='none'`
 Defaults to “saga”.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0.01, 10), “l1_ratio”: Real(0, 1)}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (`pd.DataFrame`, or `np.ndarray`) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

class `evalml.pipelines.ElasticNetRegressor` (*alpha*=0.0001, *l1_ratio*=0.15,
max_iter=1000, *normalize*=False, *random_seed*=0, ***kwargs*)

Elastic Net Regressor.

Parameters

- **alpha** (*float*) – Constant that multiplies the penalty terms. Defaults to 0.0001.
- **l1_ratio** (*float*) – The mixing parameter, with $0 \leq \text{l1_ratio} \leq 1$. Only used if `penalty='elasticnet'`. Setting `l1_ratio=0` is equivalent to using `penalty='l2'`, while setting `l1_ratio=1` is equivalent to using `penalty='l1'`. For $0 < \text{l1_ratio} < 1$, the penalty is a combination of L1 and L2. Defaults to 0.15.
- **max_iter** (*int*) – The maximum number of iterations. Defaults to 1000.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. Defaults to False.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "alpha": Real(0, 1), "l1_ratio": Real(0, 1), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Elastic Net Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.**Estimator** (*parameters=None*, *component_obj=None*, *random_seed=0*,
***kwargs*)

A component that fits and predicts given data.

To implement a new Estimator, define your own class which is a subclass of Estimator, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Estimator component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
pre-dict_uses_y	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path
<i>supported_problem_types</i>	Problem types this estimator supports

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property `feature_importance` (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type `np.ndarray`

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

property supported_problem_types (*cls*)

Problem types this estimator supports

```
class evalml.pipelines.ExtraTreesClassifier (n_estimators=100,      max_features='auto',  
                                             max_depth=6,          min_samples_split=2,  
                                             min_weight_fraction_leaf=0.0, n_jobs=- 1,  
                                             random_seed=0, **kwargs)
```

Extra Trees Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “sqrt”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “log2”, then $\text{max_features} = \text{log2}(\text{n_features})$.
 - If *None*, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.

- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.ExtraTreesRegressor (n_estimators=100,      max_features='auto',
                                           max_depth=6,          min_samples_split=2,
                                           min_weight_fraction_leaf=0.0, n_jobs=-1,
                                           random_seed=0, **kwargs)
```

Extra Trees Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_features** (*int, float or {"auto", "sqrt", "log2"}*) – The number of features to consider when looking for the best split:
 - If *int*, then consider *max_features* features at each split.
 - If *float*, then *max_features* is a fraction and $\text{int}(\text{max_features} * \text{n_features})$ features are considered at each split.
 - If “auto”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “sqrt”, then $\text{max_features} = \text{sqrt}(\text{n_features})$.
 - If “log2”, then $\text{max_features} = \text{log2}(\text{n_features})$.
 - If *None*, then $\text{max_features} = \text{n_features}$.

The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than *max_features* features. Defaults to “auto”.
- **max_depth** (*int*) – The maximum depth of the tree. Defaults to 6.
- **min_samples_split** (*int or float*) – The minimum number of samples required to split an internal node:
 - If *int*, then consider *min_samples_split* as the minimum number.
 - If *float*, then *min_samples_split* is a fraction and $\text{ceil}(\text{min_samples_split} * \text{n_samples})$ are the minimum number of samples for each split.
- **to 2.** (*Defaults*) –
- **min_weight_fraction_leaf** (*float*) – The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Defaults to 0.0.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_features”: [“auto”, “sqrt”, “log2”], “max_depth”: Integer(4, 10),}
model_family	ModelFamily.EXTRA_TREES
modifies_features	True
modifies_target	False
name	Extra Trees Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.FeatureSelector (parameters=None, component_obj=None, random_seed=0, **kwargs)
```

Selects top features based on importance weights.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>name</i>	Returns string name of this component
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that Component.default_parameters == Component().parameters.

Returns default parameters for this component.

Return type dict

describe (*self, print_name=False, return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self, X, y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an `MethodPropertyNotFoundError` exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type `pd.DataFrame`

```
class evalml.pipelines.KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, random_seed=0, **kwargs)
```

K-Nearest Neighbors Classifier.

Parameters

- **n_neighbors** (*int*) – Number of neighbors to use by default. Defaults to 5.
- **weights** (*{ 'uniform', 'distance' } or callable*) – Weight function used in prediction. Can be:
 - ‘uniform’ : uniform weights. All points in each neighborhood are weighted equally.
 - ‘distance’ : weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away.
 - [callable] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.

Defaults to “uniform”.

- **algorithm** (*{ 'auto', 'ball_tree', 'kd_tree', 'brute' }*) – Algorithm used to compute the nearest neighbors:
 - ‘ball_tree’ will use `BallTree`
 - ‘kd_tree’ will use `KDTree`
 - ‘brute’ will use a brute-force search.‘auto’ will attempt to decide the most appropriate algorithm based on the values passed to fit method. Defaults to “auto”. Note: fitting on sparse input will override the setting of this parameter, using brute force.
- **leaf_size** (*int*) – Leaf size passed to `BallTree` or `KDTree`. This can affect the speed of the construction and query, as well as the memory required to store the tree. The optimal value depends on the nature of the problem. Defaults to 30.
- **p** (*int*) – Power parameter for the Minkowski metric. When $p = 1$, this is equivalent to using `manhattan_distance` (l1), and `euclidean_distance` (l2) for $p = 2$. For arbitrary p , `minkowski_distance(l_p)` is used. Defaults to 2.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_neighbors": Integer(2, 12), "weights": ["uniform", "distance"], "algorithm": ["auto", "ball_tree", "kd_tree", "brute"], "leaf_size": Integer(10, 30), "p": Integer(1, 5), }
model_family	ModelFamily.K_NEIGHBORS
modifies_features	True
modifies_target	False
name	KNN Classifier
pre_dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's matching the input number of features as feature_importance is
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's matching the input number of features as feature_importance is not defined for KNN classifiers.

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.LightGBMClassifier(boosting_type='gbdt', learning_rate=0.1,  
                                         n_estimators=100, max_depth=0,  
                                         num_leaves=31, min_child_samples=20, bag-  
                                         ging_fraction=0.9, bagging_freq=0, n_jobs=-  
                                         1, random_seed=0, **kwargs)
```

LightGBM Classifier.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.
- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select bagging_fraction * 100 % of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “learning_rate”: Real(0.000001, 1), “boosting_type”: [“gbdt”, “dart”, “goss”, “rf”], “n_estimators”: Integer(10, 100), “max_depth”: Integer(0, 10), “num_leaves”: Integer(2, 100), “min_child_samples”: Integer(1, 100), “bagging_fraction”: Real(0.000001, 1), “bagging_freq”: Integer(0, 1),}
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Classifier
pre-dict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.LightGBMRegressor (boosting_type='gbdt', learning_rate=0.1,  
                                         n_estimators=20, max_depth=0,  
                                         num_leaves=31, min_child_samples=20, bag-  
                                         ging_fraction=0.9, bagging_freq=0, n_jobs=- 1,  
                                         random_seed=0, **kwargs)
```

LightGBM Regressor.

Parameters

- **boosting_type** (*string*) – Type of boosting to use. Defaults to “gbdt”. - ‘gbdt’ uses traditional Gradient Boosting Decision Tree - “dart”, uses Dropouts meet Multiple Additive Regression Trees - “goss”, uses Gradient-based One-Side Sampling - “rf”, uses Random Forest
- **learning_rate** (*float*) – Boosting learning rate. Defaults to 0.1.
- **n_estimators** (*int*) – Number of boosted trees to fit. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners, <=0 means no limit. Defaults to 0.

- **num_leaves** (*int*) – Maximum tree leaves for base learners. Defaults to 31.
- **min_child_samples** (*int*) – Minimum number of data needed in a child (leaf). Defaults to 20.
- **bagging_fraction** (*float*) – LightGBM will randomly select a subset of features on each iteration (tree) without resampling if this is smaller than 1.0. For example, if set to 0.8, LightGBM will select 80% of features before training each tree. This can be used to speed up training and deal with overfitting. Defaults to 0.9.
- **bagging_freq** (*int*) – Frequency for bagging. 0 means bagging is disabled. k means perform bagging at every k iteration. Every k-th iteration, LightGBM will randomly select bagging_fraction * 100 % of the data to use for the next k iterations. Defaults to 0.
- **n_jobs** (*int or None*) – Number of threads to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "learning_rate": Real(0.000001, 1), "boosting_type": ["gbdt", "dart", "goss", "rf"], "n_estimators": Integer(10, 100), "max_depth": Integer(0, 10), "num_leaves": Integer(2, 100), "min_child_samples": Integer(1, 100), "bagging_fraction": Real(0.000001, 1), "bagging_freq": Integer(0, 1), }
model_family	ModelFamily.LIGHTGBM
modifies_features	True
modifies_target	False
name	LightGBM Regressor
predict_uses_y	False
SEED_MAX	SEED_BOUNDS.max_bound
SEED_MIN	0
supported_problem_types	[ProblemTypes.REGRESSION]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class `evalml.pipelines.LinearRegressor` (*fit_intercept=True*, *normalize=False*, *n_jobs=-1*,
random_seed=0, ***kwargs*)

Linear Regressor.

Parameters

- **fit_intercept** (*boolean*) – Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered). Defaults to True.
- **normalize** (*boolean*) – If True, the regressors will be normalized before regression by subtracting the mean and dividing by the l2-norm. This parameter is ignored when `fit_intercept` is set to False. Defaults to False.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all threads. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “fit_intercept”: [True, False], “normalize”: [True, False]}
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Linear Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>feature_importance</code>	Returns importance associated with each feature.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters *file_path* (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.LogisticRegressionClassifier (penalty='l2', C=1.0,  
                                                    multi_class='auto',  
                                                    solver='lbfgs', n_jobs=-1,  
                                                    random_seed=0, **kwargs)
```

Logistic Regression Classifier.

Parameters

- **penalty** (*{ "l1", "l2", "elasticnet", "none" }*) – The norm used in penalization. Defaults to “l2”.
- **C** (*float*) – Inverse of regularization strength. Must be a positive float. Defaults to 1.0.
- **multi_class** (*{ "auto", "ovr", "multinomial" }*) – If the option chosen is “ovr”, then a binary problem is fit for each label. For “multinomial” the loss minimised is the multinomial loss fit across the entire probability distribution, even when the data is binary. “multinomial” is unavailable when solver=“liblinear”. “auto” selects “ovr” if the data is binary, or if solver=“liblinear”, and otherwise selects “multinomial”. Defaults to “auto”.
- **solver** (*{ "newton-cg", "lbfgs", "liblinear", "sag", "saga" }*) – Algorithm to use in the optimization problem. For small datasets, “liblinear” is a good choice, whereas “sag” and “saga” are faster for large ones. For multiclass problems, only

“newton-cg”, “sag”, “saga” and “lbfgs” handle multinomial loss; “liblinear” is limited to one-versus-rest schemes.

- “newton-cg”, “lbfgs”, “sag” and “saga” handle L2 or no penalty
- “liblinear” and “saga” also handle L1 penalty
- “saga” also supports “elasticnet” penalty
- “liblinear” does not support setting `penalty='none'`

Defaults to “lbfgs”.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “penalty”: [“l2”], “C”: Real(0.01, 10), }
model_family	ModelFamily.LINEAR_MODEL
modifies_features	True
modifies_target	False
name	Logistic Regression Classifier
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.MulticlassClassificationPipeline (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Pipeline subclass for all multiclass classification pipelines.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.MULTICLASS
---------------------	-------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.

continues on next page

Table 645 – continued from previous page

<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a classification model. For string and categorical targets, classes are sorted
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on objectives
<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (*PipelineBase*) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.

- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a classification model. For string and categorical targets, classes are sorted by `sorted(set(y))` and then are mapped to values between 0 and *n_classes*-1.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [*n_samples*, *n_features*]
- **y** (*pd.Series*, *np.ndarray*) – The target training labels of length [*n_samples*]

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type *dict*

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than *importance_threshold*. Defaults to zero.

Returns *plotly.Figure*, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component *inverse_transform* methods to estimator predictions in reverse order.

Components that implement *inverse_transform* are *PolynomialDetrender*, *LabelEncoder* (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns *PipelineBase* object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or *None* implies using all default values for component parameters. Defaults to *None*.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type *dict*

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape `[n_samples, n_features]`
- **objective** (*Object* or *string*) – The objective to use to make predictions
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.

- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Make probability estimates for labels.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Probability estimates

Return type *pd.DataFrame*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns *None*

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on objectives

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, or *np.ndarray*) – True labels of length [n_samples]
- **objectives** (*list*) – List of objectives to score
- **X_train** (*pd.DataFrame* or *np.ndarray*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to *None*.

Returns Transformed output.

Return type `pd.DataFrame`

```
class evalml.pipelines.OneHotEncoder(top_n=10, features_to_encode=None, categories=None, drop='if_binary', handle_unknown='ignore', handle_missing='error', random_seed=0, **kwargs)
```

A transformer that encodes categorical features in a one-hot numeric array.

Parameters

- **top_n** (*int*) – Number of categories per column to encode. If *None*, all categories will be encoded. Otherwise, the *n* most frequent will be encoded and all others will be dropped. Defaults to 10.
- **features_to_encode** (*list[str]*) – List of columns to encode. All other columns will remain untouched. If *None*, all appropriate columns will be encoded. Defaults to *None*.
- **categories** (*list*) – A two dimensional list of categories, where *categories[i]* is a list of the categories for the column at index *i*. This can also be *None*, or “*auto*” if *top_n* is not *None*. Defaults to *None*.
- **drop** (*string*, *list*) – Method (“first” or “if_binary”) to use to drop one category per feature. Can also be a list specifying which categories to drop for each feature. Defaults to “if_binary”.
- **handle_unknown** (*string*) – Whether to ignore or error for unknown categories for a feature encountered during *fit* or *transform*. If either *top_n* or *categories* is used to limit the number of categories per column, this must be “ignore”. Defaults to “ignore”.
- **handle_missing** (*string*) – Options for how to handle missing (NaN) values encountered during *fit* or *transform*. If this is set to “as_category” and NaN values are within the *n* most frequent, “nan” values will be encoded as their own column. If this is set to “error”, any missing values encountered will raise an error. Defaults to “error”.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	One Hot Encoder

Methods

<i>categories</i>	Returns a list of the unique categories to be encoded for the particular feature, in order.
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters

continues on next page

Table 646 – continued from previous page

<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the categorical features after fitting.
<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	One-hot encode the input data.

categories (*self*, *feature_name*)

Returns a list of the unique categories to be encoded for the particular feature, in order.

Parameters **feature_name** (*str*) – the name of any feature provided to one-hot encoder during fit

Returns the unique categories, in the same dtype as they were provided during fit

Return type `np.ndarray`

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type `dict`

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type `None` or `dict`

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape `[n_samples, n_features]`
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length `[n_samples]`

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on *X* and transforms *X*

Parameters

- **x** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed *X*

Return type *pd.DataFrame*

get_feature_names (*self*)

Return feature names for the categorical features after fitting.

Feature names are formatted as {column name}_{category name}. In the event of a duplicate name, an integer will be added at the end of the feature name to distinguish it.

For example, consider a dataframe with a column called “A” and category “x_y” and another column called “A_x” with “y”. In this example, the feature names would be “A_x_y” and “A_x_y_1”.

Returns The feature names after encoding, provided in the same order as *input_features*.

Return type *np.ndarray*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling *predict*, *predict_proba*, *transform*, or *feature_importances*. This can be overridden to *False* for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns *None*

transform (*self*, *X*, *y=None*)

One-hot encode the input data.

Parameters

- **x** (*pd.DataFrame*) – Features to one-hot encode.
- **y** (*pd.Series*) – Ignored.

Returns Transformed data, where each categorical feature has been encoded into numerical columns using one-hot encoding.

Return type *pd.DataFrame*

```
class evalml.pipelines.PerColumnImputer (impute_strategies=None,                de-  
                                         fault_impute_strategy='most_frequent',    ran-  
                                         dom_seed=0, **kwargs)
```

Imputes missing data according to a specified imputation strategy per column.

Parameters

- **impute_strategies** (*dict*) – Column and {"impute_strategy": strategy, "fill_value":value} pairings. Valid values for impute strategy include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to None, which uses "most_frequent" for all columns. When impute_strategy == "constant", fill_value is used to replace missing data. When None, uses 0 when imputing numerical data and "missing_value" for strings or object data types.
- **default_impute_strategy** (*str*) – Impute strategy to fall back on when none is provided for a certain column. Valid values include "mean", "median", "most_frequent", "constant" for numerical data, and "most_frequent", "constant" for object data types. Defaults to "most_frequent".
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Per Column Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputers on input data
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by imputing missing values.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputers on input data

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to fit.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by imputing missing values.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features] to transform.
- **y** (*pd.Series*, *optional*) – The target training data of length [n_samples]. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.**PipelineBase** (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Machine learning pipeline made out of transformers and an Estimator.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”].
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Build a model.
<i>get_component</i>	Returns component by name

continues on next page

Table 648 – continued from previous page

<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline’s feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

`can_tune_threshold_with_objective` (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **`pipeline`** (`PipelineBase`) – Binary classification pipeline.
- **`objective`** – Primary AutoMLSearch objective.

`clone` (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

`compute_estimator_features` (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **`X`** (`pd.DataFrame`) – Input data to the pipeline to transform.
- **`y`** (`pd.Series` or `None`) – Targets corresponding to `X`. optional.
- **`X_train`** (`pd.DataFrame` or `np.ndarray` or `None`) – Training data. Only used for time series.
- **`y_train`** (`pd.Series` or `None`) – Training labels. Only used for time series.

Returns New transformed features.

Return type `pd.DataFrame`

`static create_objectives` (*objectives*)

`property custom_name` (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type dict

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns pd.DataFrame including feature names and their corresponding importance

abstract fit (*self*, *X*, *y*)

Build a model.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training data of length [n_samples].

Returns self

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than *importance_threshold*. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters *y* (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters *file_path* (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Predicted values.

Return type pd.Series

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

abstract score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

```
class evalml.pipelines.ProphetRegressor (date_index=None, changepoint_prior_scale=0.05,
                                         seasonality_prior_scale=10,                    hol-
                                         idays_prior_scale=10,                      seasonal-
                                         ity_mode='additive',                      random_seed=0,
                                         stan_backend='CMDSTANPY', **kwargs)
```

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

More information here: <https://facebook.github.io/prophet/>

Attributes

hyper-parameter_ranges	{ "changepoint_prior_scale": Real(0.001, 0.5), "seasonality_prior_scale": Real(0.01, 10), "holidays_prior_scale": Real(0.01, 10), "seasonality_mode": ["additive", "multiplicative"], }
model_family	ModelFamily.PROPHET
modifies_features	True
modifies_target	False
name	Prophet Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.TIME_SERIES_REGRESSION]

Methods

<i>build_prophet_df</i>	
<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.
<i>fit</i>	Fits component to data
<i>get_params</i>	
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

static build_prophet_df (*X*, *y=None*, *date_column='ds'*)

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns array of 0's with len(1) as feature_importance is not defined for Prophet regressor.

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

get_params (*self*)

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X, y=None*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file

- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.**RandomForestClassifier** (*n_estimators=100*, *max_depth=6*,
n_jobs=-1, *random_seed=0*, ***kwargs*)

Random Forest Classifier.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "n_estimators": Integer(10, 1000), "max_depth": Integer(1, 10), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Classifier
predict Uses y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.**RandomForestRegressor** (*n_estimators=100*, *max_depth=6*, *n_jobs=-1*, *random_seed=0*, ***kwargs*)

Random Forest Regressor.

Parameters

- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **n_jobs** (*int* or *None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “n_estimators”: Integer(10, 1000), “max_depth”: Integer(1, 32), }
model_family	ModelFamily.RANDOM_FOREST
modifies_features	True
modifies_target	False
name	Random Forest Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data

continues on next page

Table 651 – continued from previous page

<code>load</code>	Loads component at file path
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>save</code>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.**RegressionPipeline** (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Pipeline subclass for all regression pipelines.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **custom_name** (*str*) – Custom name for the pipeline. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.REGRESSION
---------------------	-------------------------

Methods

<code>can_tune_threshold_with_objective</code>	Determine whether the threshold of a binary classification pipeline can be tuned.
<code>clone</code>	Constructs a new pipeline with the same components, parameters, and random state.
<code>compute_estimator_features</code>	Transforms the data by applying all pre-processing components.
<code>create_objectives</code>	
<code>custom_name</code>	Custom name of the pipeline.
<code>describe</code>	Outputs pipeline details including component parameters
<code>feature_importance</code>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<code>fit</code>	Build a regression model.
<code>get_component</code>	Returns component by name
<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline's feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Make predictions using selected features.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

`can_tune_threshold_with_objective` (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **`pipeline`** (`PipelineBase`) – Binary classification pipeline.
- **`objective`** – Primary AutoMLSearch objective.

`clone` (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

`compute_estimator_features` (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series* or *None*) – Targets corresponding to X. optional.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Only used for time series.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Build a regression model.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [*n_samples*, *n_features*]
- **y** (*pd.Series*, *np.ndarray*) – The target training data of length [*n_samples*]

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type *dict*

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters `filepath` (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type `graphviz.Digraph`

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters `importance_threshold` (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than `importance_threshold`. Defaults to zero.

Returns `plotly.Figure`, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component `inverse_transform` methods to estimator predictions in reverse order.

Components that implement `inverse_transform` are `PolynomialDetrender`, `LabelEncoder` (tbd).

Parameters `y` (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters `file_path` (*str*) – location to load file

Returns `PipelineBase` object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type `dict`

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape `[n_samples, n_features]`.
- **objective** (*Object* or *string*) – The objective to use to make predictions.

- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Predicted values.

Return type `pd.Series`

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)
Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns `None`

score (*self, X, y, objectives, X_train=None, y_train=None*)
Evaluate model performance on current and additional objectives

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape `[n_samples, n_features]`
- **y** (*pd.Series, or np.ndarray*) – True values of length `[n_samples]`
- **objectives** (*list*) – Non-empty list of objectives to score on
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

Returns Ordered dictionary of objective scores

Return type `dict`

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)
Transform the input.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape `[n_samples, n_features]`.
- **y** (*pd.Series*) – The target data of length `[n_samples]`. Defaults to `None`.

Returns Transformed output.

Return type `pd.DataFrame`

class `evalml.pipelines.RFClassifierSelectFromModel` (*number_features=None, n_estimators=10, max_depth=None, percent_features=0.5, threshold=-np.inf, n_jobs=-1, random_seed=0, **kwargs*)

Selects top features based on importance weights using a Random Forest classifier.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to -np.inf.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Classifier Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type list[str]

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self, X, y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series, optional*) – Target data. Ignored.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.RFRegressorSelectFromModel (number_features=None,
                                                n_estimators=10,
                                                max_depth=None, per-
                                                cent_features=0.5, threshold=-
                                                np.inf, n_jobs=-1, random_seed=0,
                                                **kwargs)
```

Selects top features based on importance weights using a Random Forest regressor.

Parameters

- **number_features** (*int*) – The maximum number of features to select. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5. Defaults to None.
- **n_estimators** (*float*) – The number of trees in the forest. Defaults to 100.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **percent_features** (*float*) – Percentage of features to use. If both percent_features and number_features are specified, take the greater number of features. Defaults to 0.5.
- **threshold** (*string or float*) – The threshold value to use for feature selection. Features whose importance is greater or equal are kept while the others are discarded. If “median”, then the threshold value is the median of the feature importances. A scaling factor (e.g., “1.25*mean”) may also be used. Defaults to -np.inf.
- **n_jobs** (*int or None*) – Number of jobs to run in parallel. -1 uses all processes. Defaults to -1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “percent_features”: Real(0.01, 1), “threshold”: [“mean”, -np.inf], }
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	RF Regressor Select From Model

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on X and transforms X
<i>get_names</i>	Get names of selected features.
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an MethodPropertyNotFoundError exception.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {“name”: name, “parameters”: parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

get_names (*self*)

Get names of selected features.

Returns List of the names of features selected

Return type *list[str]*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input data by selecting features. If the component_obj does not have a transform method, will raise an *MethodPropertyNotFoundError* exception.

Parameters

- **x** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data. Ignored.

Returns Transformed X

Return type *pd.DataFrame*

class `evalml.pipelines.SimpleImputer` (*impute_strategy='most_frequent'*, *fill_value=None*, *random_seed=0*, ***kwargs*)

Imputes missing data according to a specified imputation strategy.

Parameters

- **impute_strategy** (*string*) – Impute strategy to use. Valid values include “mean”, “median”, “most_frequent”, “constant” for numerical data, and “most_frequent”, “constant” for object data types.
- **fill_value** (*string*) – When `impute_strategy == “constant”`, `fill_value` is used to replace missing data. Defaults to 0 when imputing numerical data and “missing_value” for strings or object data types.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “impute_strategy”: [“mean”, “median”, “most_frequent”]}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Simple Imputer

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are
<i>fit_transform</i>	Fits on X and transforms X
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits imputer to data. ‘None’ values are converted to np.nan before imputation and are treated as the same.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – the input training data of shape [n_samples, n_features]
- **y** (*pd.Series*, *optional*) – the target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms input by imputing missing values. ‘None’ and np.nan values are treated as the same.

Parameters

- **X** (*pd.DataFrame*) – Data to transform
- **y** (*pd.Series, optional*) – Ignored.

Returns Transformed X

Return type pd.DataFrame

```
class evalml.pipelines.SklearnStackedEnsembleClassifier (input_pipelines=None,  
                                                         final_estimator=None,  
                                                         cv=None, n_jobs=- 1, ran-  
                                                         dom_seed=0, **kwargs)
```

Scikit-learn Stacked Ensemble Classifier.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The classifier used to combine the base estimators. If None, uses LogisticRegressionClassifier.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, if the estimator is a classifier and y is either binary or multiclass, StratifiedKFold is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.SklearnStackedEnsembleRegressor (input_pipelines=None,  
                                                    final_estimator=None,  
                                                    cv=None, n_jobs=- 1, ran-  
                                                    dom_seed=0, **kwargs)
```

Scikit-learn Stacked Ensemble Regressor.

Parameters

- **input_pipelines** (*list(PipelineBase or subclass obj)*) – List of pipeline instances to use as the base estimators. This must not be None or an empty list or else EnsembleMissingPipelinesError will be raised.
- **final_estimator** (*Estimator or subclass*) – The regressor used to combine the base estimators. If None, uses LinearRegressor.
- **cv** (*int, cross-validation generator or an iterable*) – Determines the cross-validation splitting strategy used to train final_estimator. For int/None inputs, KFold is used. Defaults to None. Possible inputs for cv are:
 - None: 3-fold cross validation
 - int: the number of folds in a (Stratified) KFold
 - An scikit-learn cross-validation generator object
 - An iterable yielding (train, test) splits
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used. Defaults to -1. - Note: there could be some multi-process errors thrown for values of *n_jobs* != 1. If this is the case, please use *n_jobs* = 1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.ENSEMBLE
modifies_features	True
modifies_target	False
name	Sklearn Stacked Ensemble Regressor
predict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 657 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for stacked ensemble classes.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Not implemented for SklearnStackedEnsembleClassifier and SklearnStackedEnsembleRegressor

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class *evalml.pipelines.StandardScaler* (*random_seed=0*, ***kwargs*)

A transformer that standardizes input features by removing the mean and scaling to unit variance.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	Standard Scaler

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>fit</i>	Fits component to data
<i>fit_transform</i>	Fits on <i>X</i> and transforms <i>X</i>
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component

continues on next page

Table 658 – continued from previous page

<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y=None*)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

class evalml.pipelines.**SVMClassifier** (*C=1.0*, *kernel='rbf'*, *gamma='auto'*, *probability=True*, *random_seed=0*, ***kwargs*)

Support Vector Machine Classifier.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses 1 / (n_features * X.var()) as value of gamma - If “auto” (default), uses 1 / n_features
- **probability** (*boolean*) – Whether to enable probability estimates. Defaults to True.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ "C": Real(0, 10), "kernel": ["poly", "rbf", "sigmoid"], "gamma": ["scale", "auto"], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Classifier
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns *self*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

class evalml.pipelines.**SVMRegressor** (*C=1.0*, *kernel='rbf'*, *gamma='auto'*, *random_seed=0*,
***kwargs*)

Support Vector Machine Regressor.

Parameters

- **C** (*float*) – The regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty. Defaults to 1.0.
- **kernel** (*{ "poly", "rbf", "sigmoid" }*) – Specifies the kernel type to be used in the algorithm. Defaults to “rbf”.
- **gamma** (*{ "scale", "auto" } or float*) – Kernel coefficient for “rbf”, “poly” and “sigmoid”. Defaults to “auto”. - If gamma=’scale’ is passed then it uses $1 / (n_features * X.var())$ as value of gamma - If “auto” (default), uses $1 / n_features$
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-parameter_ranges	{ “C”: Real(0, 10), “kernel”: [“poly”, “rbf”, “sigmoid”], “gamma”: [“scale”, “auto”], }
model_family	ModelFamily.SVM
modifies_features	True
modifies_target	False
name	SVM Regressor
pre-dict_uses_y	False
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Feature importance only works with linear kernels.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Feature importance only works with linear kernels. If the kernel isn't linear, we return a numpy array of zeros

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame*, or *np.ndarray*) – Features

Returns Probability estimates

Return type `pd.Series`

save (*self*, *file_path*, *pickle_protocol*=`cloudpickle.DEFAULT_PROTOCOL`)
Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

class `evalml.pipelines.TargetEncoder` (*cols*=`None`, *smoothing*=`1.0`, *handle_unknown*='value',
handle_missing='value', *random_seed*=`0`, ***kwargs*)

A transformer that encodes categorical features into target encodings.

Parameters

- **cols** (*list*) – Columns to encode. If `None`, all string columns will be encoded, otherwise only the columns provided will be encoded. Defaults to `None`
- **smoothing** (*float*) – The smoothing factor to apply. The larger this value is, the more influence the expected target value has on the resulting target encodings. Must be strictly larger than 0. Defaults to `1.0`
- **handle_unknown** (*string*) – Determines how to handle unknown categories for a feature encountered. Options are 'value', 'error', and 'return_nan'. Defaults to 'value', which replaces with the target mean
- **handle_missing** (*string*) – Determines how to handle missing values encountered during *fit* or *transform*. Options are 'value', 'error', and 'return_nan'. Defaults to 'value', which replaces with the target mean
- **random_seed** (*int*) – Seed for the random number generator. Defaults to `0`.

Attributes

hyper-parameter_ranges	<code>{}</code>
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
name	Target Encoder

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>get_feature_names</code>	Return feature names for the input features after fitting.

continues on next page

Table 661 – continued from previous page

<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>save</i>	Saves component at file path
<i>transform</i>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform
- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type pd.DataFrame

get_feature_names (*self*)

Return feature names for the input features after fitting.

Returns The feature names after encoding

Return type np.array

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

transform (*self*, *X*, *y*=None)

Transforms data X.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type pd.DataFrame

class evalml.pipelines.**TimeSeriesBinaryClassificationPipeline** (*component_graph*,
parameters=None, *custom_name*=None,
random_seed=0)

Pipeline base class for time series binary classification problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={“pipeline”: {“date_index”: “Date”, “max_delay”: 4, “gap”: 2}}).

- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>optimize_threshold</i>	Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Predict on future data where target is not known.
<i>predict_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>predict_proba</i>	Predict on future data where the target is unknown.
<i>predict_proba_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on current and additional objectives.

continues on next page

Table 662 – continued from previous page

<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>threshold</i>	Threshold used to make a prediction. Defaults to None.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** ([PipelineBase](#)) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (*pd.DataFrame*) – Input data to the pipeline to transform.
- **y** (*pd.Series*) – Targets corresponding to the pipeline targets.
- **X_train** (*pd.DataFrame*) – Training data used to generate features from past observations.
- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series classification pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's `__new__` method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

optimize_threshold (*self*, *X*, *y*, *y_pred_proba*, *objective*)

Optimize the pipeline threshold given the objective to use. Only used for binary problems with objectives whose thresholds can be tuned.

Parameters

- **X** (*pd.DataFrame*) – Input features
- **y** (*pd.Series*) – Input target values
- **y_pred_proba** (*pd.Series*) – The predicted probabilities of the target outputted by the pipeline
- **objective** (*ObjectiveBase*) – The objective to threshold with. Must have a tunable threshold.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self*, *X*, *y*, *X_train*, *y_train*, *objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase*, *str*, *None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type *pd.Series*

predict_proba (*self*, *X*, *X_train=None*, *y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

predict_proba_in_sample (*self*, *X_holdout*, *y_holdout*, *X_train*, *y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns *None*

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

property threshold (*self*)

Threshold used to make a prediction. Defaults to None.

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

class evalml.pipelines.**TimeSeriesClassificationPipeline** (*component_graph*, *parameters=None*, *custom_name=None*, *random_seed=0*)

Pipeline base class for time series classification problems.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={“pipeline”: {“date_index”: “Date”, “max_delay”: 4, “gap”: 2}}).
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	None
---------------------	------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Predict on future data where target is not known.
<i>predict_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>predict_proba</i>	Predict on future data where the target is unknown.
<i>predict_proba_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on current and additional objectives.
<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (`PipelineBase`) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.**compute_estimator_features** (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (`pd.DataFrame`) – Input data to the pipeline to transform.
- **y** (`pd.Series`) – Targets corresponding to the pipeline targets.
- **X_train** (`pd.DataFrame`) – Training data used to generate features from past observations.
- **y_train** (`pd.Series`) – Training targets used to generate features from past observations.

Returns New transformed features.**Return type** `pd.DataFrame`**static create_objectives** (*objectives*)**property custom_name** (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.**Returns** Dictionary of all component parameters if *return_dict* is True, else None**Return type** `dict`**property feature_importance** (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns `pd.DataFrame` including feature names and their corresponding importance**fit** (*self*, *X*, *y*)

Fit a time series classification pipeline.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – The input training data of shape `[n_samples, n_features]`.
- **y** (`pd.Series`, `np.ndarray`) – The target training targets of length `[n_samples]`.

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self, X, objective=None, X_train=None, y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object or string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame or np.ndarray or None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series or None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self, X, y, X_train, y_train, objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase, str, None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type pd.Series

predict_proba (*self, X, X_train=None, y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame or np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].

- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

predict_proba_in_sample (*self*, *X_holdout*, *y_holdout*, *X_train*, *y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type *dict*

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self*, *X*, *y=None*)

Transform the input.

Parameters

- **x** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type *pd.DataFrame*

```
class evalml.pipelines.TimeSeriesMulticlassClassificationPipeline(component_graph,
                                                                    parameters=None,
                                                                    custom_name=None,
                                                                    random_seed=0)
```

Pipeline base class for time series multiclass classification problems.

Parameters

- **component_graph** (*list* or *dict*) – List of components in order. Accepts strings or *ComponentBase* subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as *date_index*, *gap*, and *max_delay* must be specified with the “pipeline” key. For example: `Pipeline(parameters={"pipeline": {"date_index": "Date", "max_delay": 4, "gap": 2}})`.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	<code>ProblemTypes.TIME_SERIES_MULTICLASS</code>
---------------------	--

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>classes_</i>	Gets the class names for the problem.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name

continues on next page

Table 664 – continued from previous page

<code>get_hyperparameter_ranges</code>	Returns hyperparameter ranges from all components as a dictionary.
<code>graph</code>	Generate an image representing the pipeline graph.
<code>graph_feature_importance</code>	Generate a bar graph of the pipeline’s feature importance
<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline.
<code>name</code>	Name of the pipeline.
<code>new</code>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<code>parameters</code>	Parameter dictionary for this pipeline.
<code>predict</code>	Predict on future data where target is not known.
<code>predict_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>predict_proba</code>	Predict on future data where the target is unknown.
<code>predict_proba_in_sample</code>	Predict on future data where the target is known, e.g. cross validation.
<code>save</code>	Saves pipeline at file path
<code>score</code>	Evaluate model performance on current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.
<code>transform</code>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (`PipelineBase`) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

property classes_ (*self*)

Gets the class names for the problem.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (`pd.DataFrame`) – Input data to the pipeline to transform.
- **y** (`pd.Series`) – Targets corresponding to the pipeline targets.
- **X_train** (`pd.DataFrame`) – Training data used to generate generates from past observations.

- **y_train** (*pd.Series*) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type *pd.DataFrame*

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type *dict*

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns *pd.DataFrame* including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series classification pipeline.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features].
- **y** (*pd.Series*, *np.ndarray*) – The target training targets of length [n_samples].

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type *dict*

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type *graphviz.Digraph*

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns *plotly.Figure*, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.

- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self, X, y, X_train, y_train, objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].
- **objective** (*ObjectiveBase, str, None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels.

Return type *pd.Series*

predict_proba (*self, X, X_train=None, y_train=None*)

Predict on future data where the target is unknown.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

predict_proba_in_sample (*self, X_holdout, y_holdout, X_train, y_train*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features].
- **y_holdout** (*pd.Series, np.ndarray*) – Future target of shape [n_samples].
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Estimated probabilities.

Return type *pd.Series*

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None**score** (*self, X, y, objectives, X_train=None, y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].
- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.**Return type** dict**property summary** (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.**Return type** pd.DataFrame

```
class evalml.pipelines.TimeSeriesRegressionPipeline (component_graph, parameters=None, custom_name=None,
                                                    random_seed=0)
```

Pipeline base class for time series regression problems.

Parameters

- **component_graph** (*list or dict*) – List of components in order. Accepts strings or ComponentBase subclasses in the list. Note that when duplicate components are specified in a list, the duplicate component names will be modified with the component’s index in the list. For example, the component graph [Imputer, One Hot Encoder, Imputer, Logistic Regression Classifier] will have names [“Imputer”, “One Hot Encoder”, “Imputer_2”, “Logistic Regression Classifier”]
- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component’s parameters as values. An empty dictionary {} implies using all default values for component parameters. Pipeline-level parameters such as date_index, gap, and max_delay must be specified with the “pipeline” key. For example: Pipeline(parameters={“pipeline”: {“date_index”: “Date”, “max_delay”: 4, “gap”: 2}}).

- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

problem_type	ProblemTypes.TIME_SERIES_REGRESSION
---------------------	-------------------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
<i>clone</i>	Constructs a new pipeline with the same components, parameters, and random state.
<i>compute_estimator_features</i>	Transforms the data by applying all pre-processing components.
<i>create_objectives</i>	
<i>custom_name</i>	Custom name of the pipeline.
<i>describe</i>	Outputs pipeline details including component parameters
<i>feature_importance</i>	Importance associated with each feature. Features dropped by the feature selection are excluded.
<i>fit</i>	Fit a time series pipeline.
<i>get_component</i>	Returns component by name
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>graph</i>	Generate an image representing the pipeline graph.
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline's feature importance
<i>inverse_transform</i>	Apply component <i>inverse_transform</i> methods to estimator predictions in reverse order.
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline.
<i>name</i>	Name of the pipeline.
<i>new</i>	Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.
<i>parameters</i>	Parameter dictionary for this pipeline.
<i>predict</i>	Predict on future data where target is not known.
<i>predict_in_sample</i>	Predict on future data where the target is known, e.g. cross validation.
<i>save</i>	Saves pipeline at file path
<i>score</i>	Evaluate model performance on current and additional objectives.
<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
<i>transform</i>	Transform the input.

can_tune_threshold_with_objective (*self*, *objective*)

Determine whether the threshold of a binary classification pipeline can be tuned.

Parameters

- **pipeline** (`PipelineBase`) – Binary classification pipeline.
- **objective** – Primary AutoMLSearch objective.

clone (*self*)

Constructs a new pipeline with the same components, parameters, and random state.

Returns A new instance of this pipeline with identical components, parameters, and random state.

compute_estimator_features (*self*, *X*, *y=None*, *X_train=None*, *y_train=None*)

Transforms the data by applying all pre-processing components.

Parameters

- **X** (`pd.DataFrame`) – Input data to the pipeline to transform.
- **y** (`pd.Series`) – Targets corresponding to the pipeline targets.
- **X_train** (`pd.DataFrame`) – Training data used to generate features from past observations.
- **y_train** (`pd.Series`) – Training targets used to generate features from past observations.

Returns New transformed features.

Return type `pd.DataFrame`

static create_objectives (*objectives*)

property custom_name (*self*)

Custom name of the pipeline.

describe (*self*, *return_dict=False*)

Outputs pipeline details including component parameters

Parameters **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Dictionary of all component parameters if *return_dict* is True, else None

Return type `dict`

property feature_importance (*self*)

Importance associated with each feature. Features dropped by the feature selection are excluded.

Returns `pd.DataFrame` including feature names and their corresponding importance

fit (*self*, *X*, *y*)

Fit a time series pipeline.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – The input training data of shape `[n_samples, n_features]`.
- **y** (`pd.Series`, `np.ndarray`) – The target training targets of length `[n_samples]`.

Returns *self*

get_component (*self*, *name*)

Returns component by name

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

get_hyperparameter_ranges (*self*, *custom_hyperparameters*)

Returns hyperparameter ranges from all components as a dictionary.

Parameters **custom_hyperparameters** (*dict*) – Custom hyperparameters for the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the pipeline.

Return type dict

graph (*self*, *filepath=None*)

Generate an image representing the pipeline graph.

Parameters **filepath** (*str*, *optional*) – Path to where the graph should be saved. If set to None (as by default), the graph will not be saved.

Returns Graph object that can be directly displayed in Jupyter notebooks.

Return type graphviz.Digraph

graph_feature_importance (*self*, *importance_threshold=0*)

Generate a bar graph of the pipeline's feature importance

Parameters **importance_threshold** (*float*, *optional*) – If provided, graph features with a permutation importance whose absolute value is larger than importance_threshold. Defaults to zero.

Returns plotly.Figure, a bar graph showing features and their corresponding importance

inverse_transform (*self*, *y*)

Apply component inverse_transform methods to estimator predictions in reverse order.

Components that implement inverse_transform are PolynomialDetrender, LabelEncoder (tbd).

Parameters **y** (*pd.Series*) – Final component features

static load (*file_path*)

Loads pipeline at file path

Parameters **file_path** (*str*) – location to load file

Returns PipelineBase object

property model_family (*self*)

Returns model family of this pipeline.

property name (*self*)

Name of the pipeline.

new (*self*, *parameters*, *random_seed=0*)

Constructs a new instance of the pipeline with the same component graph but with a different set of parameters.

Not to be confused with python's __new__ method.

Parameters

- **parameters** (*dict*) – Dictionary with component names as keys and dictionary of that component's parameters as values. An empty dictionary or None implies using all default values for component parameters. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns A new instance of this pipeline with identical components.

property parameters (*self*)

Parameter dictionary for this pipeline.

Returns Dictionary of all component parameters.

Return type dict

predict (*self*, *X*, *objective=None*, *X_train=None*, *y_train=None*)

Predict on future data where target is not known.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features].
- **objective** (*Object* or *string*) – Used in classification problems to threshold the predictions.
- **objective** – The objective to use to make predictions.
- **X_train** (*pd.DataFrame* or *np.ndarray* or *None*) – Training data. Ignored. Only used for time series.
- **y_train** (*pd.Series* or *None*) – Training labels. Ignored. Only used for time series.

predict_in_sample (*self*, *X*, *y*, *X_train*, *y_train*, *objective=None*)

Predict on future data where the target is known, e.g. cross validation.

Parameters

- **X_holdout** (*pd.DataFrame* or *np.ndarray*) – Future data of shape [n_samples, n_features]
- **y_holdout** (*pd.Series*, *np.ndarray*) – Future target of shape [n_samples]
- **X_train** (*pd.DataFrame*, *np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features]
- **y_train** (*pd.Series*, *np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train]
- **objective** (*ObjectiveBase*, *str*, *None*) – Objective used to threshold predicted probabilities, optional.

Returns Estimated labels

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves pipeline at file path

Parameters

- **file_path** (*str*) – location to save file
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score (*self*, *X*, *y*, *objectives*, *X_train=None*, *y_train=None*)

Evaluate model performance on current and additional objectives.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – True labels of length [n_samples].

- **objectives** (*list*) – Non-empty list of objectives to score on.
- **X_train** (*pd.DataFrame, np.ndarray*) – Data the pipeline was trained on of shape [n_samples_train, n_features].
- **y_train** (*pd.Series, np.ndarray*) – Targets used to train the pipeline of shape [n_samples_train].

Returns Ordered dictionary of objective scores.

Return type dict

property summary (*self*)

A short summary of the pipeline structure, describing the list of components used. Example: Logistic Regression Classifier w/ Simple Imputer + One Hot Encoder

transform (*self, X, y=None*)

Transform the input.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features].
- **y** (*pd.Series*) – The target data of length [n_samples]. Defaults to None.

Returns Transformed output.

Return type pd.DataFrame

class evalml.pipelines.**Transformer** (*parameters=None, component_obj=None, random_seed=0, **kwargs*)

A component that may or may not need fitting that transforms data. These components are used before an estimator.

To implement a new Transformer, define your own class which is a subclass of Transformer, including a name and a list of acceptable ranges for any parameters to be tuned during the automl search (hyperparameters). Define an `__init__` method which sets up any necessary state and objects. Make sure your `__init__` only uses standard keyword arguments and calls `super().__init__()` with a parameters dict. You may also override the `fit`, `transform`, `fit_transform` and other methods in this class if appropriate.

To see some examples, check out the definitions of any Transformer component.

Parameters

- **parameters** (*dict*) – Dictionary of parameters for the component. Defaults to None.
- **component_obj** (*obj*) – Third-party objects useful in component implementation. Defaults to None.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False

Methods

<code>clone</code>	Constructs a new component with the same parameters and random state.
<code>default_parameters</code>	Returns the default parameters for this component.
<code>describe</code>	Describe a component and its parameters
<code>fit</code>	Fits component to data
<code>fit_transform</code>	Fits on X and transforms X
<code>load</code>	Loads component at file path
<code>name</code>	Returns string name of this component
<code>needs_fitting</code>	Returns boolean determining if component needs fitting before
<code>parameters</code>	Returns the parameters which were used to initialize the component
<code>save</code>	Saves component at file path
<code>transform</code>	Transforms data X.

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

fit_transform (*self*, *X*, *y=None*)

Fits on X and transforms X

Parameters

- **X** (*pd.DataFrame*) – Data to fit and transform

- **y** (*pd.Series*) – Target data

Returns Transformed X

Return type *pd.DataFrame*

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns *ComponentBase* object

property name (*cls*)

Returns string name of this component

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling `predict`, `predict_proba`, `transform`, or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

transform (*self*, *X*, *y=None*)

Transforms data *X*.

Parameters

- **X** (*pd.DataFrame*) – Data to transform.
- **y** (*pd.Series*, *optional*) – Target data.

Returns Transformed X

Return type *pd.DataFrame*

```
class evalml.pipelines.XGBoostClassifier(eta=0.1, max_depth=6, min_child_weight=1,  
                                         n_estimators=100, random_seed=0,  
                                         eval_metric='logloss', n_jobs=12, **kwargs)
```

XGBoost Classifier.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ “eta”: Real(0.000001, 1), “max_depth”: Integer(1, 10), “min_child_weight”: Real(1, 10), “n_estimators”: Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Classifier
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_BINARY, ProblemTypes.TIME_SERIES_MULTICLASS,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before
<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool, optional*) – whether to print name of component
- **return_dict** (*bool, optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self, X, y=None*)

Fits component to data

Parameters

- **X** (*list, pd.DataFrame or np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list, pd.Series, np.ndarray, optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform, or feature_importances. This can be overridden to False for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame, np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type pd.Series

predict_proba (*self, X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame, or np.ndarray*) – Features

Returns Probability estimates

Return type pd.Series

save (*self, file_path, pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns None

```
class evalml.pipelines.XGBoostRegressor(eta=0.1, max_depth=6, min_child_weight=1,
                                         n_estimators=100, random_seed=0, n_jobs=12,
                                         **kwargs)
```

XGBoost Regressor.

Parameters

- **eta** (*float*) – Boosting learning rate. Defaults to 0.1.
- **max_depth** (*int*) – Maximum tree depth for base learners. Defaults to 6.
- **min_child_weight** (*float*) – Minimum sum of instance weight (hessian) needed in a child. Defaults to 1.0
- **n_estimators** (*int*) – Number of gradient boosted trees. Equivalent to number of boosting rounds. Defaults to 100.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int*) – Number of parallel threads used to run xgboost. Note that creating thread contention will significantly slow down the algorithm. Defaults to 12.

Attributes

hyper-parameter_ranges	{ "eta": Real(0.000001, 1), "max_depth": Integer(1, 20), "min_child_weight": Real(1, 10), "n_estimators": Integer(1, 1000), }
model_family	ModelFamily.XGBOOST
modifies_features	True
modifies_target	False
name	XGBoost Regressor
pre-dict_uses_y	False
SEED_MAX	None
SEED_MIN	None
supported_problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION,]

Methods

<i>clone</i>	Constructs a new component with the same parameters and random state.
<i>default_parameters</i>	Returns the default parameters for this component.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Returns importance associated with each feature.
<i>fit</i>	Fits component to data
<i>load</i>	Loads component at file path
<i>needs_fitting</i>	Returns boolean determining if component needs fitting before

continues on next page

Table 668 – continued from previous page

<i>parameters</i>	Returns the parameters which were used to initialize the component
<i>predict</i>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels.
<i>save</i>	Saves component at file path

clone (*self*)

Constructs a new component with the same parameters and random state.

Returns A new instance of this component with identical parameters and random state.

default_parameters (*cls*)

Returns the default parameters for this component.

Our convention is that `Component.default_parameters == Component().parameters`.

Returns default parameters for this component.

Return type dict

describe (*self*, *print_name=False*, *return_dict=False*)

Describe a component and its parameters

Parameters

- **print_name** (*bool*, *optional*) – whether to print name of component
- **return_dict** (*bool*, *optional*) – whether to return description as dictionary in the format {"name": name, "parameters": parameters}

Returns prints and returns dictionary

Return type None or dict

property feature_importance (*self*)

Returns importance associated with each feature.

Returns Importance associated with each feature

Return type np.ndarray

fit (*self*, *X*, *y=None*)

Fits component to data

Parameters

- **X** (*list*, *pd.DataFrame* or *np.ndarray*) – The input training data of shape [n_samples, n_features]
- **y** (*list*, *pd.Series*, *np.ndarray*, *optional*) – The target training data of length [n_samples]

Returns self

static load (*file_path*)

Loads component at file path

Parameters **file_path** (*str*) – Location to load file

Returns ComponentBase object

needs_fitting (*self*)

Returns boolean determining if component needs fitting before calling predict, predict_proba, transform,

or `feature_importances`. This can be overridden to `False` for components that do not need to be fit or whose fit methods do nothing.

property parameters (*self*)

Returns the parameters which were used to initialize the component

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*, *np.ndarray*) – Data of shape [n_samples, n_features]

Returns Predicted values

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame*, *or np.ndarray*) – Features

Returns Probability estimates

Return type *pd.Series*

save (*self*, *file_path*, *pickle_protocol=cloudpickle.DEFAULT_PROTOCOL*)

Saves component at file path

Parameters

- **file_path** (*str*) – Location to save file
- **pickle_protocol** (*int*) – The pickle data stream format.

Returns `None`

Preprocessing

Subpackages

data_splitters

Submodules

balanced_classification_sampler

Module Contents

Classes Summary

BalancedClassificationSampler

Class for balanced classification downsampler.

Contents

`class evalml.preprocessing.data_splitters.balanced_classification_sampler.BalancedClassifi`

Class for balanced classification downsampler.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. Overrides sampling_ratio if provided. Defaults to None.
- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below min_percentage. Must be greater than 0. Defaults to 100.
- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above min_samples. To determine severe imbalance, the minority class must have a class ratio below this and must occur less often than min_samples. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<code>fit_resample</code>	Resampling technique for this sampler.
---------------------------	--

fit_resample (*self*, *X*, *y*)

Resampling technique for this sampler.

Parameters

- **X** (*pd.DataFrame*) – Training data to fit and resample
- **y** (*pd.Series*) – Training data targets to fit and resample

Returns Indices to keep for training data

Return type list

sampler_base

Module Contents

Classes Summary

<i>SamplerBase</i>	Base class for all custom samplers.
--------------------	-------------------------------------

Contents

class evalml.preprocessing.data_splitters.sampler_base.**SamplerBase** (*random_seed=0*)
Base class for all custom samplers.

Parameters **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<i>fit_resample</i>	Resample the input data with this sampling strategy.
---------------------	--

abstract fit_resample (*self, X, y*)
Resample the input data with this sampling strategy.

Parameters

- **X** (*pd.DataFrame*) – Training data to fit and resample.
- **y** (*pd.Series*) – Training data targets to fit and resample.

Returns resampled X and y data for oversampling or indices to keep for undersampling.

Return type Tuple(pd.DataFrame, pd.Series) or list

time_series_split

Module Contents

Classes Summary

<i>TimeSeriesSplit</i>	Rolling Origin Cross Validation for time series problems.
------------------------	---

Contents

class evalml.preprocessing.data_splitters.time_series_split.**TimeSeriesSplit** (*max_delay=0*,
gap=0,
date_index=None,
n_splits=3)

Rolling Origin Cross Validation for time series problems.

This class uses `max_delay` and `gap` values to take into account that evalml time series pipelines perform some feature and target engineering, e.g delaying input features and shifting the target variable by the desired amount. If the data that will be split already has all the features and appropriate target values, and then set `max_delay` and `gap` to 0.

Parameters

- **max_delay** (*int*) – Max delay value for feature engineering. Time series pipelines create delayed features from existing features. This process will introduce NaNs into the first `max_delay` number of rows. The splitter uses the last `max_delay` number of rows from the previous split as the first `max_delay` number of rows of the current split to avoid “throwing out” more data than in necessary. Defaults to 0.
- **gap** (*int*) – Gap used in time series problem. Time series pipelines shift the target variable by `gap` rows. Defaults to 0.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Defaults to None.
- **n_splits** (*int*) – number of data splits to make. Defaults to 3.

Methods

<code>get_n_splits</code>	Get the number of data splits.
<code>split</code>	Get the time series splits.

get_n_splits (*self*, *X=None*, *y=None*, *groups=None*)
Get the number of data splits.

split (*self*, *X*, *y=None*, *groups=None*)
Get the time series splits.

X and *y* are assumed to be sorted in ascending time order. This method can handle passing in empty or None *X* and *y* data but note that *X* and *y* cannot be None or empty at the same time.

Parameters

- **X** (*pd.DataFrame*, *None*) – Features to split.
- **y** (*pd.DataFrame*, *None*) – Target variable to split.
- **groups** – Ignored but kept for compatibility with sklearn api.

Returns Iterator of (train, test) indices tuples.

training_validation_split

Module Contents

Classes Summary

<i>TrainingValidationSplit</i>	Split the training data into training and validation sets.
--------------------------------	--

Contents

class evalml.preprocessing.data_splitters.training_validation_split.**TrainingValidationSplit**

Split the training data into training and validation sets.

Parameters

- **test_size** (*float*) – What percentage of data points should be included in the validation set. Defaults to the complement of *train_size* if *train_size* is set, and 0.25 otherwise.
- **train_size** (*float*) – What percentage of data points should be included in the training set. Defaults to the complement of *test_size*
- **shuffle** (*boolean*) – Whether to shuffle the data before splitting. Defaults to False.
- **stratify** (*list*) – Splits the data in a stratified fashion, using this argument as class labels. Defaults to None.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<i>get_n_splits</i>	Returns the number of splits of this object
<i>split</i>	Divides the data into training and testing sets

static **get_n_splits** ()

Returns the number of splits of this object

split (*self*, *X*, *y=None*)

Divides the data into training and testing sets

Parameters

- **X** (*pd.DataFrame*) – Dataframe of points to split
- **y** (*pd.Series*) – Series of points to split

Returns Indices to split data into training and test set

Return type list

Package Contents

Classes Summary

<code>BalancedClassificationSampler</code>	Class for balanced classification downsampler.
<code>SamplerBase</code>	Base class for all custom samplers.
<code>TimeSeriesSplit</code>	Rolling Origin Cross Validation for time series problems.
<code>TrainingValidationSplit</code>	Split the training data into training and validation sets.

Contents

class evalml.preprocessing.data_splitters.**BalancedClassificationSampler** (*sampling_ratio=0.25, sampling_ratio_dict=None, min_samples=100, min_percentage=0.1, random_seed=0*)

Class for balanced classification downsampler.

Parameters

- **sampling_ratio** (*float*) – The smallest minority:majority ratio that is accepted as ‘balanced’. For instance, a 1:4 ratio would be represented as 0.25, while a 1:1 ratio is 1.0. Must be between 0 and 1, inclusive. Defaults to 0.25.
- **sampling_ratio_dict** (*dict*) – A dictionary specifying the desired balanced ratio for each target value. Overrides `sampling_ratio` if provided. Defaults to None.
- **min_samples** (*int*) – The minimum number of samples that we must have for any class, pre or post sampling. If a class must be downsampled, it will not be downsampled past this value. To determine severe imbalance, the minority class must occur less often than this and must have a class ratio below `min_percentage`. Must be greater than 0. Defaults to 100.
- **min_percentage** (*float*) – The minimum percentage of the minimum class to total dataset that we tolerate, as long as it is above `min_samples`. To determine severe imbalance, the minority class must have a class ratio below this and must occur less often than `min_samples`. Must be between 0 and 0.5, inclusive. Defaults to 0.1.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<code>fit_resample</code>	Resampling technique for this sampler.
---------------------------	--

fit_resample (*self, X, y*)
Resampling technique for this sampler.

Parameters

- **X** (*pd.DataFrame*) – Training data to fit and resample
- **y** (*pd.Series*) – Training data targets to fit and resample

Returns Indices to keep for training data

Return type list

class evalml.preprocessing.data_splitters.**SamplerBase** (*random_seed=0*)
Base class for all custom samplers.

Parameters **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<i>fit_resample</i>	Resample the input data with this sampling strategy.
---------------------	--

abstract fit_resample (*self, X, y*)
Resample the input data with this sampling strategy.

Parameters

- **X** (*pd.DataFrame*) – Training data to fit and resample.
- **y** (*pd.Series*) – Training data targets to fit and resample.

Returns resampled X and y data for oversampling or indices to keep for undersampling.

Return type Tuple(pd.DataFrame, pd.Series) or list

class evalml.preprocessing.data_splitters.**TimeSeriesSplit** (*max_delay=0, gap=0, date_index=None, n_splits=3*)

Rolling Origin Cross Validation for time series problems.

This class uses max_delay and gap values to take into account that evalml time series pipelines perform some feature and target engineering, e.g delaying input features and shifting the target variable by the desired amount. If the data that will be split already has all the features and appropriate target values, and then set max_delay and gap to 0.

Parameters

- **max_delay** (*int*) – Max delay value for feature engineering. Time series pipelines create delayed features from existing features. This process will introduce NaNs into the first max_delay number of rows. The splitter uses the last max_delay number of rows from the previous split as the first max_delay number of rows of the current split to avoid “throwing out” more data than is necessary. Defaults to 0.
- **gap** (*int*) – Gap used in time series problem. Time series pipelines shift the target variable by gap rows. Defaults to 0.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Defaults to None.
- **n_splits** (*int*) – number of data splits to make. Defaults to 3.

Methods

<i>get_n_splits</i>	Get the number of data splits.
<i>split</i>	Get the time series splits.

get_n_splits (*self, X=None, y=None, groups=None*)
Get the number of data splits.

split (*self, X, y=None, groups=None*)
Get the time series splits.

X and y are assumed to be sorted in ascending time order. This method can handle passing in empty or

None X and y data but note that X and y cannot be None or empty at the same time.

Parameters

- **x** (*pd.DataFrame*, *None*) – Features to split.
- **y** (*pd.DataFrame*, *None*) – Target variable to split.
- **groups** – Ignored but kept for compatibility with sklearn api.

Returns Iterator of (train, test) indices tuples.

```
class evalml.preprocessing.data_splitters.TrainingValidationSplit (test_size=None,  
                                                                train_size=None,  
                                                                shuf-  
                                                                fle=False,  
                                                                strat-  
                                                                ify=None,  
                                                                ran-  
                                                                dom_seed=0)
```

Split the training data into training and validation sets.

Parameters

- **test_size** (*float*) – What percentage of data points should be included in the validation set. Defaults to the complement of *train_size* if *train_size* is set, and 0.25 otherwise.
- **train_size** (*float*) – What percentage of data points should be included in the training set. Defaults to the complement of *test_size*
- **shuffle** (*boolean*) – Whether to shuffle the data before splitting. Defaults to False.
- **stratify** (*list*) – Splits the data in a stratified fashion, using this argument as class labels. Defaults to None.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<code>get_n_splits</code>	Returns the number of splits of this object
<code>split</code>	Divides the data into training and testing sets

static `get_n_splits()`

Returns the number of splits of this object

split (*self*, *X*, *y=None*)

Divides the data into training and testing sets

Parameters

- **x** (*pd.DataFrame*) – Dataframe of points to split
- **y** (*pd.Series*) – Series of points to split

Returns Indices to split data into training and test set

Return type list

Submodules

utils

Module Contents

Functions

<code>load_data</code>	Load features and target from file.
<code>number_of_features</code>	Get the number of features of each specific dtype in a DataFrame.
<code>split_data</code>	Splits data into train and test sets.
<code>target_distribution</code>	Get the target distributions.

Contents

`evalml.preprocessing.utils.load_data` (*path*, *index*, *target*, *n_rows=None*, *drop=None*, *verbose=True*, ***kwargs*)

Load features and target from file.

Parameters

- **path** (*str*) – Path to file or a http/ftp/s3 URL
- **index** (*str*) – Column for index
- **target** (*str*) – Column for target
- **n_rows** (*int*) – Number of rows to return
- **drop** (*list*) – List of columns to drop
- **verbose** (*bool*) – If True, prints information about features and target

Returns Features matrix and target

Return type `pd.DataFrame`, `pd.Series`

`evalml.preprocessing.utils.number_of_features` (*dtypes*)

Get the number of features of each specific dtype in a DataFrame.

Parameters **dtypes** (`pd.Series`) – DataFrame.dtypes to get the number of features for

Returns dtypes and the number of features for each input type

Return type `pd.Series`

`evalml.preprocessing.utils.split_data` (*X*, *y*, *problem_type*, *problem_configuration=None*, *test_size=0.2*, *random_seed=0*)

Splits data into train and test sets.

Parameters

- **X** (`pd.DataFrame` or `np.ndarray`) – data of shape `[n_samples, n_features]`
- **y** (`pd.Series`, or `np.ndarray`) – target data of length `[n_samples]`
- **problem_type** (*str* or `ProblemTypes`) – type of supervised learning problem. see `evalml.problem_types.problemtype.all_problem_types` for a full list.

- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, and `max_delay` variables.
- **test_size** (*float*) – What percentage of data points should be included in the test set. Defaults to 0.2 (20%).
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Feature and target data each split into train and test sets

Return type `pd.DataFrame`, `pd.DataFrame`, `pd.Series`, `pd.Series`

`evalml.preprocessing.utils.target_distribution` (*targets*)

Get the target distributions.

Parameters **targets** (*pd.Series*) – Target data

Returns Target data and their frequency distribution as percentages.

Return type `pd.Series`

Package Contents

Classes Summary

<code>TimeSeriesSplit</code>	Rolling Origin Cross Validation for time series problems.
<code>TrainingValidationSplit</code>	Split the training data into training and validation sets.

Functions

<code>load_data</code>	Load features and target from file.
<code>number_of_features</code>	Get the number of features of each specific dtype in a <code>DataFrame</code> .
<code>split_data</code>	Splits data into train and test sets.
<code>target_distribution</code>	Get the target distributions.

Contents

`evalml.preprocessing.load_data` (*path*, *index*, *target*, *n_rows=None*, *drop=None*, *verbose=True*, ***kwargs*)

Load features and target from file.

Parameters

- **path** (*str*) – Path to file or a `http/ftp/s3` URL
- **index** (*str*) – Column for index
- **target** (*str*) – Column for target
- **n_rows** (*int*) – Number of rows to return
- **drop** (*list*) – List of columns to drop

- **verbose** (*bool*) – If True, prints information about features and target

Returns Features matrix and target

Return type `pd.DataFrame`, `pd.Series`

`evalml.preprocessing.number_of_features` (*dtypes*)

Get the number of features of each specific dtype in a DataFrame.

Parameters **dtypes** (*pd.Series*) – `DataFrame.dtypes` to get the number of features for

Returns dtypes and the number of features for each input type

Return type `pd.Series`

`evalml.preprocessing.split_data` (*X*, *y*, *problem_type*, *problem_configuration=None*,
test_size=0.2, *random_seed=0*)

Splits data into train and test sets.

Parameters

- **X** (*pd.DataFrame* or *np.ndarray*) – data of shape `[n_samples, n_features]`
- **y** (*pd.Series*, or *np.ndarray*) – target data of length `[n_samples]`
- **problem_type** (*str* or *ProblemTypes*) – type of supervised learning problem. see `evalml.problem_types.problemtype.all_problem_types` for a full list.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, and `max_delay` variables.
- **test_size** (*float*) – What percentage of data points should be included in the test set. Defaults to 0.2 (20%).
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Returns Feature and target data each split into train and test sets

Return type `pd.DataFrame`, `pd.DataFrame`, `pd.Series`, `pd.Series`

`evalml.preprocessing.target_distribution` (*targets*)

Get the target distributions.

Parameters **targets** (*pd.Series*) – Target data

Returns Target data and their frequency distribution as percentages.

Return type `pd.Series`

class `evalml.preprocessing.TimeSeriesSplit` (*max_delay=0*, *gap=0*, *date_index=None*,
n_splits=3)

Rolling Origin Cross Validation for time series problems.

This class uses `max_delay` and `gap` values to take into account that evalml time series pipelines perform some feature and target engineering, e.g delaying input features and shifting the target variable by the desired amount. If the data that will be split already has all the features and appropriate target values, and then set `max_delay` and `gap` to 0.

Parameters

- **max_delay** (*int*) – Max delay value for feature engineering. Time series pipelines create delayed features from existing features. This process will introduce NaNs into the first `max_delay` number of rows. The splitter uses the last `max_delay` number of rows from the previous split as the first `max_delay` number of rows of the current split to avoid “throwing out” more data than is necessary. Defaults to 0.

- **gap** (*int*) – Gap used in time series problem. Time series pipelines shift the target variable by gap rows. Defaults to 0.
- **date_index** (*str*) – Name of the column containing the datetime information used to order the data. Defaults to None.
- **n_splits** (*int*) – number of data splits to make. Defaults to 3.

Methods

<code>get_n_splits</code>	Get the number of data splits.
<code>split</code>	Get the time series splits.

get_n_splits (*self*, *X=None*, *y=None*, *groups=None*)

Get the number of data splits.

split (*self*, *X*, *y=None*, *groups=None*)

Get the time series splits.

X and *y* are assumed to be sorted in ascending time order. This method can handle passing in empty or None *X* and *y* data but note that *X* and *y* cannot be None or empty at the same time.

Parameters

- **x** (*pd.DataFrame*, *None*) – Features to split.
- **y** (*pd.DataFrame*, *None*) – Target variable to split.
- **groups** – Ignored but kept for compatibility with sklearn api.

Returns Iterator of (train, test) indices tuples.

class evalml.preprocessing.**TrainingValidationSplit** (*test_size=None*, *train_size=None*, *shuffle=False*, *stratify=None*, *random_seed=0*)

Split the training data into training and validation sets.

Parameters

- **test_size** (*float*) – What percentage of data points should be included in the validation set. Defaults to the complement of *train_size* if *train_size* is set, and 0.25 otherwise.
- **train_size** (*float*) – What percentage of data points should be included in the training set. Defaults to the complement of *test_size*
- **shuffle** (*boolean*) – Whether to shuffle the data before splitting. Defaults to False.
- **stratify** (*list*) – Splits the data in a stratified fashion, using this argument as class labels. Defaults to None.
- **random_seed** (*int*) – The seed to use for random sampling. Defaults to 0.

Methods

<code>get_n_splits</code>	Returns the number of splits of this object
<code>split</code>	Divides the data into training and testing sets

static get_n_splits ()

Returns the number of splits of this object

split (*self*, *X*, *y=None*)

Divides the data into training and testing sets

Parameters

- **x** (*pd.DataFrame*) – Dataframe of points to split
- **y** (*pd.Series*) – Series of points to split

Returns Indices to split data into training and test set

Return type list

Problem Types**Submodules****problem_types****Module Contents****Classes Summary**

<i>ProblemTypes</i>	Enum defining the supported types of machine learning problems.
---------------------	---

Contents

class evalml.problem_types.problem_types.**ProblemTypes**
Enum defining the supported types of machine learning problems.

Attributes

BINARY	Binary classification problem.
MULTI-CLASS	Multiclass classification problem.
REGRESSION	Regression problem.
TIME_SERIES_BINARY	Time series binary classification problem.
TIME_SERIES_MULTICLASS	Time series multiclass classification problem.
TIME_SERIES_REGRESSION	Time series regression problem.

Methods

<i>all_problem_types</i>	Get a list of all defined problem types.
<i>name</i>	The name of the Enum member.
<i>value</i>	The value of the Enum member.

all_problem_types (*cls*)
Get a list of all defined problem types.

Returns list

Return type list(*ProblemTypes*)

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

utils

Module Contents

Functions

<code>detect_problem_type</code>	Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)
<code>handle_problem_types</code>	Handles problem_type by either returning the ProblemTypes or converting from a str.
<code>is_binary</code>	Determines if the provided problem_type is a binary classification problem type
<code>is_classification</code>	Determines if the provided problem_type is a classification problem type
<code>is_multiclass</code>	Determines if the provided problem_type is a multiclass classification problem type
<code>is_regression</code>	Determines if the provided problem_type is a regression problem type
<code>is_time_series</code>	Determines if the provided problem_type is a time series problem type

Contents

`evalml.problem_types.utils.detect_problem_type(y)`

Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)

Ignores missing and null data

Parameters `y` (*pd.Series*) – the target labels to predict

Returns ProblemType Enum

Return type ProblemType

Example

```
>>> y = pd.Series([0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1])
>>> problem_type = detect_problem_type(y)
>>> assert problem_type == ProblemTypes.BINARY
```

`evalml.problem_types.utils.handle_problem_types(problem_type)`

Handles problem_type by either returning the ProblemTypes or converting from a str.

Parameters `problem_type` (*str or ProblemTypes*) – Problem type that needs to be handled

Returns ProblemTypes

`evalml.problem_types.utils.is_binary(problem_type)`

Determines if the provided problem_type is a binary classification problem type

Parameters `problem_type` (*str or ProblemTypes*) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided problem_type is a binary classification problem type.

Return type bool

`evalml.problem_types.utils.is_classification(problem_type)`

Determines if the provided problem_type is a classification problem type

Parameters `problem_type` (*str or ProblemTypes*) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided problem_type is a classification problem type.

Return type bool

`evalml.problem_types.utils.is_multiclass(problem_type)`

Determines if the provided problem_type is a multiclass classification problem type

Parameters `problem_type` (*str or ProblemTypes*) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided problem_type is a multiclass classification problem type.

Return type bool

`evalml.problem_types.utils.is_regression(problem_type)`

Determines if the provided problem_type is a regression problem type

Parameters `problem_type` (*str or ProblemTypes*) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided problem_type is a regression problem type.

Return type bool

`evalml.problem_types.utils.is_time_series(problem_type)`

Determines if the provided problem_type is a time series problem type

Parameters `problem_type` (*str or ProblemTypes*) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided problem_type is a time series problem type.

Return type bool

Package Contents

Classes Summary

<i>ProblemTypes</i>	Enum defining the supported types of machine learning problems.
---------------------	---

Functions

<code>detect_problem_type</code>	Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)
<code>handle_problem_types</code>	Handles <code>problem_type</code> by either returning the <code>ProblemTypes</code> or converting from a str.
<code>is_binary</code>	Determines if the provided <code>problem_type</code> is a binary classification problem type
<code>is_classification</code>	Determines if the provided <code>problem_type</code> is a classification problem type
<code>is_multiclass</code>	Determines if the provided <code>problem_type</code> is a multiclass classification problem type
<code>is_regression</code>	Determines if the provided <code>problem_type</code> is a regression problem type
<code>is_time_series</code>	Determines if the provided <code>problem_type</code> is a time series problem type

Contents

`evalml.problem_types.detect_problem_type(y)`

Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)

Ignores missing and null data

Parameters `y` (`pd.Series`) – the target labels to predict

Returns `ProblemType` Enum

Return type `ProblemType`

Example

```
>>> y = pd.Series([0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1])
>>> problem_type = detect_problem_type(y)
>>> assert problem_type == ProblemTypes.BINARY
```

`evalml.problem_types.handle_problem_types(problem_type)`

Handles `problem_type` by either returning the `ProblemTypes` or converting from a str.

Parameters `problem_type` (`str` or `ProblemTypes`) – Problem type that needs to be handled

Returns `ProblemTypes`

`evalml.problem_types.is_binary(problem_type)`

Determines if the provided `problem_type` is a binary classification problem type

Parameters `problem_type` (`str` or `ProblemTypes`) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided `problem_type` is a binary classification problem type.

Return type `bool`

`evalml.problem_types.is_classification(problem_type)`

Determines if the provided `problem_type` is a classification problem type

Parameters `problem_type` (*str* or `ProblemTypes`) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided `problem_type` is a classification problem type.

Return type bool

`evalml.problem_types.is_multiclass(problem_type)`

Determines if the provided `problem_type` is a multiclass classification problem type

Parameters `problem_type` (*str* or `ProblemTypes`) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided `problem_type` is a multiclass classification problem type.

Return type bool

`evalml.problem_types.is_regression(problem_type)`

Determines if the provided `problem_type` is a regression problem type

Parameters `problem_type` (*str* or `ProblemTypes`) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided `problem_type` is a regression problem type.

Return type bool

`evalml.problem_types.is_time_series(problem_type)`

Determines if the provided `problem_type` is a time series problem type

Parameters `problem_type` (*str* or `ProblemTypes`) – type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.

Returns Whether or not the provided `problem_type` is a time series problem type.

Return type bool

class `evalml.problem_types.ProblemTypes`

Enum defining the supported types of machine learning problems.

Attributes

BINARY	Binary classification problem.
MULTI-CLASS	Multiclass classification problem.
REGRESSION	Regression problem.
TIME_SERIES_BINARY	Time series binary classification problem.
TIME_SERIES_MULTICLASS	Time series multiclass classification problem.
TIME_SERIES_REGRESSION	Time series regression problem.

Methods

<code>all_problem_types</code>	Get a list of all defined problem types.
<code>name</code>	The name of the Enum member.
<code>value</code>	The value of the Enum member.

`all_problem_types` (*cls*)

Get a list of all defined problem types.

Returns list

Return type list(*ProblemTypes*)

name (*self*)

The name of the Enum member.

value (*self*)

The value of the Enum member.

Tuners

Submodules

grid_search_tuner

Module Contents

Classes Summary

GridSearchTuner

Grid Search Optimizer, which generates all of the possible points to search for using a grid.

Contents

class evalml.tuners.grid_search_tuner.**GridSearchTuner** (*pipeline_hyperparameter_ranges*,
n_points=10, ran-
dom_seed=0)

Grid Search Optimizer, which generates all of the possible points to search for using a grid.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges corresponding to a pipeline’s parameters
- **n_points** (*int*) – The number of points to sample from along each dimension defined in the *space* argument. Defaults to 10.
- **random_seed** (*int*) – Seed for random number generator. Unused in this class, defaults to 0.

Example

```
>>> tuner = GridSearchTuner({'My Component': {'param a': [0.0, 10.0], 'param b': [
↪ 'a', 'b', 'c']}}, n_points=5)
>>> proposal = tuner.propose()
>>> assert proposal.keys() == {'My Component'}
>>> assert proposal['My Component'] == {'param a': 0.0, 'param b': 'a'}
```

Methods

<code>add</code>	Not applicable to grid search tuner as generated parameters are
<code>is_search_space_exhausted</code>	Checks if it is possible to generate a set of valid parameters. Stores generated parameters in
<code>propose</code>	Returns parameters from <code>_grid_points</code> iterations

add (*self*, *pipeline_parameters*, *score*)

Not applicable to grid search tuner as generated parameters are not dependent on scores of previous parameters.

Parameters

- **pipeline_parameters** (*dict*) – a dict of the parameters used to evaluate a pipeline
- **score** (*float*) – the score obtained by evaluating the pipeline with the provided parameters

is_search_space_exhausted (*self*)

Checks if it is possible to generate a set of valid parameters. Stores generated parameters in `self.curr_params` to be returned by `propose()`.

Raises `NoParamsException` – If a search space is exhausted, then this exception is thrown.

Returns If no more valid parameters exists in the search space, return false.

Return type bool

propose (*self*)

Returns parameters from `_grid_points` iterations

If all possible combinations of parameters have been scored, then `NoParamsException` is raised.

Returns proposed pipeline parameters

Return type dict

random_search_tuner

Module Contents

Classes Summary

<code>RandomSearchTuner</code>	Random Search Optimizer.
--------------------------------	--------------------------

Contents

class `evalml.tuners.random_search_tuner.RandomSearchTuner` (*pipeline_hyperparameter_ranges*, *with_replacement=False*, *replacement_max_attempts=10*, *random_seed=0*)

Random Search Optimizer.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges cor-

responding to a pipeline's parameters

- **with_replacement** (*bool*) – If false, only unique hyperparameters will be shown
- **replacement_max_attempts** (*int*) – The maximum number of tries to get a unique set of random parameters. Only used if tuner is initialized with `with_replacement=True`
- **random_seed** (*int*) – Seed for random number generator. Defaults to 0.

Example

```
>>> tuner = RandomSearchTuner({'My Component': {'param a': [0.0, 10.0], 'param b':
↳ ['a', 'b', 'c']}}, random_seed=42)
>>> proposal = tuner.propose()
>>> assert proposal.keys() == {'My Component'}
>>> assert proposal['My Component'] == {'param a': 3.7454011884736254, 'param b':
↳ 'c'}
```

Methods

<code>add</code>	Not applicable to random search tuner as generated parameters are
<code>is_search_space_exhausted</code>	Checks if it is possible to generate a set of valid parameters. Stores generated parameters in
<code>propose</code>	Generate a unique set of parameters.

add (*self*, *pipeline_parameters*, *score*)

Not applicable to random search tuner as generated parameters are not dependent on scores of previous parameters.

Parameters

- **pipeline_parameters** (*dict*) – A dict of the parameters used to evaluate a pipeline
- **score** (*float*) – The score obtained by evaluating the pipeline with the provided parameters

is_search_space_exhausted (*self*)

Checks if it is possible to generate a set of valid parameters. Stores generated parameters in `self.curr_params` to be returned by `propose()`.

Raises **NoParamsException** – If a search space is exhausted, then this exception is thrown.

Returns If no more valid parameters exists in the search space, return false.

Return type bool

propose (*self*)

Generate a unique set of parameters.

If tuner was initialized with `with_replacement=True` and the tuner is unable to generate a unique set of parameters after `replacement_max_attempts` tries, then `NoParamsException` is raised.

Returns Proposed pipeline parameters

Return type dict

skopt_tuner

Module Contents

Classes Summary

<i>SKOptTuner</i>	Bayesian Optimizer.
-------------------	---------------------

Attributes Summary

<i>logger</i>

Contents

evalml.tuners.skopt_tuner.logger

class evalml.tuners.skopt_tuner.**SKOptTuner** (*pipeline_hyperparameter_ranges*, *random_seed=0*)
Bayesian Optimizer.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – A set of hyperparameter ranges corresponding to a pipeline’s parameters.
- **random_seed** (*int*) – The seed for the random number generator. Defaults to 0.

Methods

<i>add</i>	Add score to sample.
<i>is_search_space_exhausted</i>	Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.
<i>propose</i>	Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

add (*self*, *pipeline_parameters*, *score*)
Add score to sample.

Parameters

- **pipeline_parameters** (*dict*) – A dict of the parameters used to evaluate a pipeline
- **score** (*float*) – The score obtained by evaluating the pipeline with the provided parameters

Returns None

is_search_space_exhausted (*self*)

Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.

Returns Returns true if all possible parameters in a search space has been scored.

Return type bool

propose (*self*)

Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

Returns Proposed pipeline parameters

Return type dict

tuner

Module Contents

Classes Summary

<i>Tuner</i>	Defines API for base Tuner classes.
--------------	-------------------------------------

Contents

class evalml.tuners.tuner.**Tuner** (*pipeline_hyperparameter_ranges*, *random_seed=0*)

Defines API for base Tuner classes.

Tuners implement different strategies for sampling from a search space. They're used in EvalML to search the space of pipeline hyperparameters.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges corresponding to a pipeline's parameters.
- **random_seed** (*int*) – The random state. Defaults to 0.

Methods

<i>add</i>	Register a set of hyperparameters with the score obtained from training a pipeline with those hyperparameters.
<i>is_search_space_exhausted</i>	Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.
<i>propose</i>	Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

abstract add (*self*, *pipeline_parameters*, *score*)

Register a set of hyperparameters with the score obtained from training a pipeline with those hyperparameters.

Parameters

- **pipeline_parameters** (*dict*) – a dict of the parameters used to evaluate a pipeline

- **score** (*float*) – the score obtained by evaluating the pipeline with the provided parameters

Returns None

is_search_space_exhausted (*self*)

Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.

Returns Returns true if all possible parameters in a search space has been scored.

Return type bool

abstract propose (*self*)

Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

Returns Proposed pipeline parameters

Return type dict

tuner_exceptions

Module Contents

Contents

exception evalml.tuners.tuner_exceptions.NoParamsException

Raised when a tuner exhausts its search space and runs out of parameters to propose.

exception evalml.tuners.tuner_exceptions.ParameterError

Raised when a tuner encounters an error with the parameters being used with it.

Package Contents

Classes Summary

<i>GridSearchTuner</i>	Grid Search Optimizer, which generates all of the possible points to search for using a grid.
<i>RandomSearchTuner</i>	Random Search Optimizer.
<i>SKOptTuner</i>	Bayesian Optimizer.
<i>Tuner</i>	Defines API for base Tuner classes.

Exceptions Summary

Contents

class evalml.tuners.**GridSearchTuner** (*pipeline_hyperparameter_ranges*, *n_points=10*, *random_seed=0*)

Grid Search Optimizer, which generates all of the possible points to search for using a grid.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges corresponding to a pipeline’s parameters
- **n_points** (*int*) – The number of points to sample from along each dimension defined in the *space* argument. Defaults to 10.
- **random_seed** (*int*) – Seed for random number generator. Unused in this class, defaults to 0.

Example

```
>>> tuner = GridSearchTuner({'My Component': {'param a': [0.0, 10.0], 'param b': [
↪ 'a', 'b', 'c']}}, n_points=5)
>>> proposal = tuner.propose()
>>> assert proposal.keys() == {'My Component'}
>>> assert proposal['My Component'] == {'param a': 0.0, 'param b': 'a'}
```

Methods

<code>add</code>	Not applicable to grid search tuner as generated parameters are
<code>is_search_space_exhausted</code>	Checks if it is possible to generate a set of valid parameters. Stores generated parameters in
<code>propose</code>	Returns parameters from <code>_grid_points</code> iterations

add (*self*, *pipeline_parameters*, *score*)

Not applicable to grid search tuner as generated parameters are not dependent on scores of previous parameters.

Parameters

- **pipeline_parameters** (*dict*) – a dict of the parameters used to evaluate a pipeline
- **score** (*float*) – the score obtained by evaluating the pipeline with the provided parameters

is_search_space_exhausted (*self*)

Checks if it is possible to generate a set of valid parameters. Stores generated parameters in `self.curr_params` to be returned by `propose()`.

Raises `NoParamsException` – If a search space is exhausted, then this exception is thrown.

Returns If no more valid parameters exists in the search space, return false.

Return type bool

propose (*self*)

Returns parameters from `_grid_points` iterations

If all possible combinations of parameters have been scored, then `NoParamsException` is raised.

Returns proposed pipeline parameters

Return type dict

exception evalml.tuners.NoParamsException

Raised when a tuner exhausts its search space and runs out of parameters to propose.

exception evalml.tuners.ParameterError

Raised when a tuner encounters an error with the parameters being used with it.

class evalml.tuners.RandomSearchTuner (pipeline_hyperparameter_ranges, with_replacement=False, replacement_max_attempts=10, random_seed=0)

Random Search Optimizer.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges corresponding to a pipeline’s parameters
- **with_replacement** (*bool*) – If false, only unique hyperparameters will be shown
- **replacement_max_attempts** (*int*) – The maximum number of tries to get a unique set of random parameters. Only used if tuner is initialized with with_replacement=True
- **random_seed** (*int*) – Seed for random number generator. Defaults to 0.

Example

```
>>> tuner = RandomSearchTuner({'My Component': {'param a': [0.0, 10.0], 'param b':
↳ ['a', 'b', 'c']}}, random_seed=42)
>>> proposal = tuner.propose()
>>> assert proposal.keys() == {'My Component'}
>>> assert proposal['My Component'] == {'param a': 3.7454011884736254, 'param b':
↳ 'c'}
```

Methods

<i>add</i>	Not applicable to random search tuner as generated parameters are
<i>is_search_space_exhausted</i>	Checks if it is possible to generate a set of valid parameters. Stores generated parameters in
<i>propose</i>	Generate a unique set of parameters.

add (*self*, pipeline_parameters, score)

Not applicable to random search tuner as generated parameters are not dependent on scores of previous parameters.

Parameters

- **pipeline_parameters** (*dict*) – A dict of the parameters used to evaluate a pipeline
- **score** (*float*) – The score obtained by evaluating the pipeline with the provided parameters

is_search_space_exhausted (*self*)

Checks if it is possible to generate a set of valid parameters. Stores generated parameters in *self*. *curr_params* to be returned by *propose()*.

Raises *NoParamsException* – If a search space is exhausted, then this exception is thrown.

Returns If no more valid parameters exists in the search space, return false.

Return type bool

propose (*self*)

Generate a unique set of parameters.

If tuner was initialized with `with_replacement=True` and the tuner is unable to generate a unique set of parameters after `replacement_max_attempts` tries, then `NoParamsException` is raised.

Returns Proposed pipeline parameters

Return type dict

class evalml.tuners.**SKOptTuner** (*pipeline_hyperparameter_ranges*, *random_seed=0*)

Bayesian Optimizer.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – A set of hyperparameter ranges corresponding to a pipeline’s parameters.
- **random_seed** (*int*) – The seed for the random number generator. Defaults to 0.

Methods

<i>add</i>	Add score to sample.
<i>is_search_space_exhausted</i>	Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.
<i>propose</i>	Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

add (*self*, *pipeline_parameters*, *score*)

Add score to sample.

Parameters

- **pipeline_parameters** (*dict*) – A dict of the parameters used to evaluate a pipeline
- **score** (*float*) – The score obtained by evaluating the pipeline with the provided parameters

Returns None

is_search_space_exhausted (*self*)

Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.

Returns Returns true if all possible parameters in a search space has been scored.

Return type bool

propose (*self*)

Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

Returns Proposed pipeline parameters

Return type dict

class evalml.tuners.**Tuner** (*pipeline_hyperparameter_ranges*, *random_seed=0*)

Defines API for base Tuner classes.

Tuners implement different strategies for sampling from a search space. They're used in EvalML to search the space of pipeline hyperparameters.

Parameters

- **pipeline_hyperparameter_ranges** (*dict*) – a set of hyperparameter ranges corresponding to a pipeline's parameters.
- **random_seed** (*int*) – The random state. Defaults to 0.

Methods

<i>add</i>	Register a set of hyperparameters with the score obtained from training a pipeline with those hyperparameters.
<i>is_search_space_exhausted</i>	Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.
<i>propose</i>	Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

abstract add (*self*, *pipeline_parameters*, *score*)

Register a set of hyperparameters with the score obtained from training a pipeline with those hyperparameters.

Parameters

- **pipeline_parameters** (*dict*) – a dict of the parameters used to evaluate a pipeline
- **score** (*float*) – the score obtained by evaluating the pipeline with the provided parameters

Returns None

is_search_space_exhausted (*self*)

Optional. If possible search space for tuner is finite, this method indicates whether or not all possible parameters have been scored.

Returns Returns true if all possible parameters in a search space has been scored.

Return type bool

abstract propose (*self*)

Returns a suggested set of parameters to train and score a pipeline with, based off the search space dimensions and prior samples.

Returns Proposed pipeline parameters

Return type dict

Utils

Submodules

base_meta

Module Contents

Classes Summary

<i>BaseMeta</i>	Metaclass that overrides creating a new component or pipeline by wrapping methods with validators and setters
-----------------	---

Contents

class evalml.utils.base_meta.**BaseMeta**

Metaclass that overrides creating a new component or pipeline by wrapping methods with validators and setters

Attributes

FIT_METHODS	['fit', 'fit_transform']
METHODS_TO_CHECK	['predict', 'predict_proba', 'transform', 'inverse_transform']
PROPERTIES_TO_CHECK	['feature_importance']

Methods

<i>register</i>	Register a virtual subclass of an ABC.
<i>set_fit</i>	

register (*cls*, *subclass*)

Register a virtual subclass of an ABC.

Returns the subclass, to allow usage as a class decorator.

classmethod **set_fit** (*cls*, *method*)

cli_utils

Module Contents

Functions

<i>get_evalml_root</i>	Gets location where evalml is installed.
------------------------	--

continues on next page

Table 710 – continued from previous page

<code>get_installed_packages</code>	Get dictionary mapping installed package names to their versions.
<code>get_sys_info</code>	Returns system information.
<code>print_deps</code>	Prints the version number of each dependency.
<code>print_info</code>	Prints information about the system, evalml, and dependencies of evalml.
<code>print_sys_info</code>	Prints system information.

Contents

`evalml.utils.cli_utils.get_evalml_root()`

Gets location where evalml is installed.

Returns Location where evalml is installed.

`evalml.utils.cli_utils.get_installed_packages()`

Get dictionary mapping installed package names to their versions.

Returns Dictionary mapping installed package names to their versions.

`evalml.utils.cli_utils.get_sys_info()`

Returns system information.

Returns List of tuples about system stats.

`evalml.utils.cli_utils.print_deps()`

Prints the version number of each dependency.

Returns None

`evalml.utils.cli_utils.print_info()`

Prints information about the system, evalml, and dependencies of evalml.

Returns None

`evalml.utils.cli_utils.print_sys_info()`

Prints system information.

Returns None

gen_utils

Module Contents

Classes Summary

<code>classproperty</code>	Allows function to be accessed as a class level property.
----------------------------	---

Functions

<code>convert_to_seconds</code>	Converts a string describing a length of time to its length in seconds.
<code>deprecate_arg</code>	Helper to raise warnings when a deprecated arg is used.
<code>drop_rows_with_nans</code>	Drop rows that have any NaNs in all dataframes or series.
<code>get_importable_subclasses</code>	Get importable subclasses of a base class. Used to list all of our
<code>get_random_seed</code>	Given a <code>numpy.random.RandomState</code> object, generate an int representing a seed value for another random number generator. Or, if given an int, return that int.
<code>get_random_state</code>	Generates a <code>numpy.random.RandomState</code> instance using seed.
<code>import_or_raise</code>	Attempts to import the requested library by name.
<code>is_all_numeric</code>	Checks if the given DataFrame contains only numeric values
<code>jupyter_check</code>	Get whether or not the code is being run in a Jupyter environment (such as Jupyter Notebook or Jupyter Lab)
<code>pad_with_nans</code>	Pad the beginning num_to_pad rows with nans.
<code>safe_repr</code>	Convert the given value into a string that can safely be used for repr
<code>save_plot</code>	Saves fig to filepath if specified, or to a default location if not.

Attributes Summary

<code>logger</code>
<code>SEED_BOUNDS</code>

Contents

class `evalml.utils.gen_utils.classproperty` (*func*)

Allows function to be accessed as a class level property.

Example:

```
class LogisticRegressionBinaryPipeline(PipelineBase):
    component_graph = ['Simple Imputer', 'Logistic Regression Classifier']

    @classproperty
    def summary(cls):
        summary = ""
        for component in cls.component_graph:
            component = handle_component_class(component)
            summary += component.name + " + "
        return summary
```

(continues on next page)

(continued from previous page)

```

assert LogisticRegressionBinaryPipeline.summary == "Simple Imputer + Logistic_
↪Regression Classifier + "
assert LogisticRegressionBinaryPipeline().summary == "Simple Imputer + Logistic_
↪Regression Classifier + "

```

`evalml.utils.gen_utils.convert_to_seconds(input_str)`
 Converts a string describing a length of time to its length in seconds.

`evalml.utils.gen_utils.deprecate_arg(old_arg, new_arg, old_value, new_value)`
 Helper to raise warnings when a deprecated arg is used.

Parameters

- **old_arg** (*str*) – Name of old/deprecated argument.
- **new_arg** (*str*) – Name of new argument.
- **old_value** (*Any*) – Value the user passed in for the old argument.
- **new_value** (*Any*) – Value the user passed in for the new argument.

Returns old_value if not None, else new_value

`evalml.utils.gen_utils.drop_rows_with_nans(*pd_data)`
 Drop rows that have any NaNs in all dataframes or series.

Parameters ***pd_data** (*sequence of pd.Series or pd.DataFrame or None*) –

Returns list of pd.DataFrame or pd.Series or None

`evalml.utils.gen_utils.get_importable_subclasses(base_class, used_in_automl=True)`
 Get importable subclasses of a base class. Used to list all of our estimators, transformers, components and pipelines dynamically.

Parameters

- **base_class** (*abc.ABCMeta*) – Base class to find all of the subclasses for.
- **args** (*list*) – Args used to instantiate the subclass. [{}] for a pipeline, and [] for all other classes.
- **used_in_automl** – Not all components/pipelines/estimators are used in automl search. If True, only include those subclasses that are used in the search. This would mean excluding classes related to ExtraTrees, ElasticNet, and Baseline estimators.

Returns List of subclasses.

`evalml.utils.gen_utils.get_random_seed(random_state, min_bound=SEED_BOUNDS.min_bound, max_bound=SEED_BOUNDS.max_bound)`
 Given a `numpy.random.RandomState` object, generate an int representing a seed value for another random number generator. Or, if given an int, return that int.

To protect against invalid input to a particular library's random number generator, if an int value is provided, and it is outside the bounds "[min_bound, max_bound)", the value will be projected into the range between the min_bound (inclusive) and max_bound (exclusive) using modular arithmetic.

Parameters

- **random_state** (*int, numpy.random.RandomState*) – random state
- **min_bound** (*None, int*) – if not default of None, will be min bound when generating seed (inclusive). Must be less than max_bound.

- **max_bound** (*None, int*) – if not default of *None*, will be max bound when generating seed (exclusive). Must be greater than *min_bound*.

Returns seed for random number generator

Return type *int*

`evalml.utils.gen_utils.get_random_state(seed)`

Generates a `numpy.random.RandomState` instance using *seed*.

Parameters *seed* (*None, int, np.random.RandomState object*) – *seed* to use to generate `numpy.random.RandomState`. Must be between `SEED_BOUNDS.min_bound` and `SEED_BOUNDS.max_bound`, inclusive. Otherwise, an exception will be thrown.

`evalml.utils.gen_utils.import_or_raise(library, error_msg=None, warning=False)`

Attempts to import the requested library by name. If the import fails, raises an `ImportError` or warning.

Parameters

- **library** (*str*) – the name of the library
- **error_msg** (*str*) – error message to return if the import fails
- **warning** (*bool*) – if *True*, `import_or_raise` gives a warning instead of `ImportError`. Defaults to *False*.

`evalml.utils.gen_utils.is_all_numeric(df)`

Checks if the given `DataFrame` contains only numeric values

Parameters *df* (*pd.DataFrame*) – The `DataFrame` to check data types of.

Returns *True* if all the columns are numeric and are not missing any values, *False* otherwise.

`evalml.utils.gen_utils.jupyter_check()`

Get whether or not the code is being run in a `IPython` environment (such as `Jupyter Notebook` or `Jupyter Lab`)

Parameters *None* –

Returns *True* if `IPython`, *False* otherwise

Return type *Boolean*

`evalml.utils.gen_utils.logger`

`evalml.utils.gen_utils.pad_with_nans(pd_data, num_to_pad)`

Pad the beginning *num_to_pad* rows with nans.

Parameters *pd_data* (*pd.DataFrame or pd.Series*) – Data to pad.

Returns *pd.DataFrame or pd.Series*

`evalml.utils.gen_utils.safe_repr(value)`

Convert the given value into a string that can safely be used for repr

Parameters *value* – the item to convert

Returns String representation of the value

`evalml.utils.gen_utils.save_plot(fig, filepath=None, format='png', interactive=False, return_filepath=False)`

Saves *fig* to *filepath* if specified, or to a default location if not.

Parameters

- **fig** (*Figure*) – Figure to be saved.
- **filepath** (*str or Path, optional*) – Location to save file. Default is with file-name “test_plot”.

- **format** (*str*) – Extension for figure to be saved as. Ignored if interactive is True and fig
- **of type** `plotly.Figure`. Defaults to 'png'. (*is*) –
- **interactive** (*bool*, *optional*) – If True and fig is of type `plotly.Figure`, saves the fig as interactive
- **of static** (*instead*) –
- **format will be set to 'html'**. Defaults to **False**. (*and*) –
- **return_filepath** (*bool*, *optional*) – Whether to return the final filepath the image is saved to. Defaults to False.

Returns String representing the final filepath the image was saved to if `return_filepath` is set to True. Defaults to None.

`evalml.utils.gen_utils.SEED_BOUNDS`

logger

Module Contents

Functions

<code>get_logger</code>	
<code>log_subtitle</code>	
<code>log_title</code>	
<code>time_elapsed</code>	How much time has elapsed since the search started.

Contents

`evalml.utils.logger.get_logger` (*name*)

`evalml.utils.logger.log_subtitle` (*logger*, *title*, *underline*='=')

`evalml.utils.logger.log_title` (*logger*, *title*)

`evalml.utils.logger.time_elapsed` (*start_time*)

How much time has elapsed since the search started.

Parameters `start_time` (*int*) – Time when search started.

Returns elapsed time formatted as a string [H:]MM:SS

Return type `str`

update_checker

Module Contents

Contents

`evalml.utils.update_checker.method`

woodwork_utils

Module Contents

Functions

<code>infer_feature_types</code>	Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.
----------------------------------	--

Attributes Summary

<code>numeric_and_boolean_ww</code>

Contents

`evalml.utils.woodwork_utils.infer_feature_types` (*data*, *feature_types=None*)

Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.
If a column's type is not specified, it will be inferred by Woodwork.

Parameters

- **data** (*pd.DataFrame*, *pd.Series*) – Input data to convert to a Woodwork data structure.
- **feature_types** (*string*, *ww.logical_type obj*, *dict*, *optional*) – If data is a 2D structure, feature_types must be a dictionary mapping column names to the type of data represented in the column. If data is a 1D structure, then feature_types must be a Woodwork logical type or a string representing a Woodwork logical type (“Double”, “Integer”, “Boolean”, “Categorical”, “Datetime”, “NaturalLanguage”)

Returns A Woodwork data structure where the data type of each column was either specified or inferred.

`evalml.utils.woodwork_utils.numeric_and_boolean_ww`

Package Contents

Classes Summary

<code>classproperty</code>	Allows function to be accessed as a class level property.
----------------------------	---

Functions

<code>convert_to_seconds</code>	Converts a string describing a length of time to its length in seconds.
<code>deprecate_arg</code>	Helper to raise warnings when a deprecated arg is used.
<code>drop_rows_with_nans</code>	Drop rows that have any NaNs in all dataframes or series.
<code>get_importable_subclasses</code>	Get importable subclasses of a base class. Used to list all of our
<code>get_logger</code>	
<code>get_random_seed</code>	Given a <code>numpy.random.RandomState</code> object, generate an int representing a seed value for another random number generator. Or, if given an int, return that int.
<code>get_random_state</code>	Generates a <code>numpy.random.RandomState</code> instance using seed.
<code>import_or_raise</code>	Attempts to import the requested library by name.
<code>infer_feature_types</code>	Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.
<code>is_all_numeric</code>	Checks if the given DataFrame contains only numeric values
<code>jupyter_check</code>	Get whether or not the code is being run in a Jupyter environment (such as Jupyter Notebook or Jupyter Lab)
<code>log_subtitle</code>	
<code>log_title</code>	
<code>pad_with_nans</code>	Pad the beginning <code>num_to_pad</code> rows with nans.
<code>safe_repr</code>	Convert the given value into a string that can safely be used for repr
<code>save_plot</code>	Saves fig to filepath if specified, or to a default location if not.

Attributes Summary

<code>SEED_BOUNDS</code>	
--------------------------	--

Contents

class evalml.utils.**classproperty** (*func*)

Allows function to be accessed as a class level property.

Example:

```
class LogisticRegressionBinaryPipeline(PipelineBase):
    component_graph = ['Simple Imputer', 'Logistic Regression Classifier']

    @classproperty
    def summary(cls):
        summary = ""
        for component in cls.component_graph:
            component = handle_component_class(component)
            summary += component.name + " + "
        return summary

assert LogisticRegressionBinaryPipeline.summary == "Simple Imputer + Logistic_
↪Regression Classifier + "
assert LogisticRegressionBinaryPipeline().summary == "Simple Imputer + Logistic_
↪Regression Classifier + "
```

evalml.utils.**convert_to_seconds** (*input_str*)

Converts a string describing a length of time to its length in seconds.

evalml.utils.**deprecate_arg** (*old_arg, new_arg, old_value, new_value*)

Helper to raise warnings when a deprecated arg is used.

Parameters

- **old_arg** (*str*) – Name of old/deprecated argument.
- **new_arg** (*str*) – Name of new argument.
- **old_value** (*Any*) – Value the user passed in for the old argument.
- **new_value** (*Any*) – Value the user passed in for the new argument.

Returns old_value if not None, else new_value

evalml.utils.**drop_rows_with_nans** (**pd_data*)

Drop rows that have any NaNs in all dataframes or series.

Parameters **pd_data* (*sequence of pd.Series or pd.DataFrame or None*) –

Returns list of pd.DataFrame or pd.Series or None

evalml.utils.**get_importable_subclasses** (*base_class, used_in_automl=True*)

Get importable subclasses of a base class. Used to list all of our estimators, transformers, components and pipelines dynamically.

Parameters

- **base_class** (*abc.ABCMeta*) – Base class to find all of the subclasses for.
- **args** (*list*) – Args used to instantiate the subclass. [{}] for a pipeline, and [] for all other classes.
- **used_in_automl** – Not all components/pipelines/estimators are used in automl search. If True, only include those subclasses that are used in the search. This would mean excluding classes related to ExtraTrees, ElasticNet, and Baseline estimators.

Returns List of subclasses.

`evalml.utils.get_logger(name)`

`evalml.utils.get_random_seed(random_state, min_bound=SEED_BOUNDS.min_bound, max_bound=SEED_BOUNDS.max_bound)`

Given a `numpy.random.RandomState` object, generate an `int` representing a seed value for another random number generator. Or, if given an `int`, return that `int`.

To protect against invalid input to a particular library's random number generator, if an `int` value is provided, and it is outside the bounds "[`min_bound`, `max_bound`)", the value will be projected into the range between the `min_bound` (inclusive) and `max_bound` (exclusive) using modular arithmetic.

Parameters

- **random_state** (*int*, *numpy.random.RandomState*) – random state
- **min_bound** (*None*, *int*) – if not default of *None*, will be min bound when generating seed (inclusive). Must be less than `max_bound`.
- **max_bound** (*None*, *int*) – if not default of *None*, will be max bound when generating seed (exclusive). Must be greater than `min_bound`.

Returns seed for random number generator

Return type `int`

`evalml.utils.get_random_state(seed)`

Generates a `numpy.random.RandomState` instance using seed.

Parameters **seed** (*None*, *int*, *np.random.RandomState object*) – seed to use to generate `numpy.random.RandomState`. Must be between `SEED_BOUNDS.min_bound` and `SEED_BOUNDS.max_bound`, inclusive. Otherwise, an exception will be thrown.

`evalml.utils.import_or_raise(library, error_msg=None, warning=False)`

Attempts to import the requested library by name. If the import fails, raises an `ImportError` or warning.

Parameters

- **library** (*str*) – the name of the library
- **error_msg** (*str*) – error message to return if the import fails
- **warning** (*bool*) – if *True*, `import_or_raise` gives a warning instead of `ImportError`. Defaults to *False*.

`evalml.utils.infer_feature_types(data, feature_types=None)`

Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.

If a column's type is not specified, it will be inferred by Woodwork.

Parameters

- **data** (*pd.DataFrame*, *pd.Series*) – Input data to convert to a Woodwork data structure.
- **feature_types** (*string*, *ww.logical_type obj*, *dict*, *optional*) – If data is a 2D structure, `feature_types` must be a dictionary mapping column names to the type of data represented in the column. If data is a 1D structure, then `feature_types` must be a Woodwork logical type or a string representing a Woodwork logical type ("Double", "Integer", "Boolean", "Categorical", "Datetime", "NaturalLanguage")

Returns A Woodwork data structure where the data type of each column was either specified or inferred.

`evalml.utils.is_all_numeric(df)`

Checks if the given DataFrame contains only numeric values

Parameters `df` (*pd.DataFrame*) – The DataFrame to check data types of.

Returns True if all the columns are numeric and are not missing any values, False otherwise.

`evalml.utils.jupyter_check()`

Get whether or not the code is being run in a Ipython environment (such as Jupyter Notebook or Jupyter Lab)

Parameters None –

Returns True if Ipython, False otherwise

Return type Boolean

`evalml.utils.log_subtitle(logger, title, underline='')`

`evalml.utils.log_title(logger, title)`

`evalml.utils.pad_with_nans(pd_data, num_to_pad)`

Pad the beginning num_to_pad rows with nans.

Parameters `pd_data` (*pd.DataFrame* or *pd.Series*) – Data to pad.

Returns *pd.DataFrame* or *pd.Series*

`evalml.utils.safe_repr(value)`

Convert the given value into a string that can safely be used for repr

Parameters `value` – the item to convert

Returns String representation of the value

`evalml.utils.save_plot(fig, filepath=None, format='png', interactive=False, return_filepath=False)`

Saves fig to filepath if specified, or to a default location if not.

Parameters

- **fig** (*Figure*) – Figure to be saved.
- **filepath** (*str* or *Path*, *optional*) – Location to save file. Default is with filename “test_plot”.
- **format** (*str*) – Extension for figure to be saved as. Ignored if interactive is True and fig
- **of type** `plotly.Figure`. Defaults to 'png'. (*is*) –
- **interactive** (*bool*, *optional*) – If True and fig is of type `plotly.Figure`, saves the fig as interactive
- **of static** (*instead*) –
- **format will be set to 'html'. Defaults to False.** (*and*) –
- **return_filepath** (*bool*, *optional*) – Whether to return the final filepath the image is saved to. Defaults to False.

Returns String representing the final filepath the image was saved to if `return_filepath` is set to True. Defaults to None.

`evalml.utils.SEED_BOUNDS`

Package Contents

Classes Summary

<i>AutoMLSearch</i>	Automated Pipeline search.
---------------------	----------------------------

Functions

<i>search</i>	Given data and configuration, run an automl search.
<i>search_iterative</i>	Given data and configuration, run an automl search.

Contents

```
class evalml.AutoMLSearch(X_train=None, y_train=None, problem_type=None, objective='auto',
                           max_iterations=None, max_time=None, patience=None, tolerance=None, data_splitter=None,
                           allowed_component_graphs=None, allowed_model_families=None, start_iteration_callback=None,
                           add_result_callback=None, error_callback=None, additional_objectives=None,
                           alternate_thresholding_objective='F1', random_seed=0, n_jobs=-1, tuner_class=None,
                           optimize_thresholds=True, ensembling=False, max_batches=None, problem_configuration=None,
                           train_best_pipeline=True, pipeline_parameters=None, custom_hyperparameters=None,
                           sampler_method='auto', sampler_balanced_ratio=0.25, _ensembling_split_size=0.2,
                           _pipelines_per_batch=5, _automl_algorithm='iterative', engine='sequential', verbose=False)
```

Automated Pipeline search.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.
- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to 'auto', chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.
- **max_iterations** (*int*) – Maximum number of iterations to search. If `max_iterations` and `max_time` is not set, then `max_iterations` will default to `max_iterations` of 5.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in

seconds. For strings, time can be specified as seconds, minutes, or hours.

- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If None, early stopping is disabled. Defaults to None.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if patience is not None. Defaults to None.
- **allowed_component_graphs** (*dict*) – A dictionary of lists or ComponentGraphs indicating the component graphs allowed in the search. The format should follow { “Name_0”: [list_of_components], “Name_1”: [ComponentGraph(...)] }

The default of None indicates all pipeline component graphs for this problem type are allowed. Setting this field will cause allowed_model_families to be ignored.

e.g. allowed_component_graphs = { “My_Graph”: [“Imputer”, “One Hot Encoder”, “Random Forest Classifier”] }

- **allowed_model_families** (*list(str, ModelFamily)*) – The model families to search. The default of None searches over all model families. Run `evalml.pipelines.components.utils.allowed_model_families(“binary”)` to see options. Change *binary* to *multiclass* or *regression* depending on the problem type. Note that if allowed_pipelines is provided, this parameter will be ignored.
- **data_splitter** (*sklearn.model_selection.BaseCrossValidator*) – Data splitting method to use. Defaults to StratifiedKFold.
- **tuner_class** – The tuner class to use. Defaults to SKOptTuner.
- **optimize_thresholds** (*bool*) – Whether or not to optimize the binary pipeline threshold. Defaults to True.
- **start_iteration_callback** (*callable*) – Function called before each pipeline training iteration. Callback function takes three positional parameters: The pipeline instance and the AutoMLSearch object.
- **add_result_callback** (*callable*) – Function called after each pipeline training iteration. Callback function takes three positional parameters: A dictionary containing the training results for the new pipeline, an untrained_pipeline containing the parameters used during training, and the AutoMLSearch object.
- **error_callback** (*callable*) – Function called when *search()* errors and raises an Exception. Callback function takes three positional parameters: the Exception raised, the traceback, and the AutoMLSearch object. Must also accepts kwargs, so AutoMLSearch is able to pass along other appropriate parameters by default. Defaults to None, which will call *log_error_callback*.
- **additional_objectives** (*list*) – Custom set of objectives to score on. Will override default objectives for problem type if not empty.
- **alternate_thresholding_objective** (*str*) – The objective to use for thresholding binary classification pipelines if the main objective provided isn’t tuneable. Defaults to F1.
- **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.
- **n_jobs** (*int or None*) – Non-negative integer describing level of parallelism used for pipelines. None and 1 are equivalent. If set to -1, all CPUs are used. For n_jobs below -1, (n_cpus + 1 + n_jobs) are used.

- **ensembling** (*boolean*) – If True, runs ensembling in a separate batch after every allowed pipeline class has been iterated over. If the number of unique pipelines to search over per batch is one, ensembling will not run. Defaults to False.
- **max_batches** (*int*) – The maximum number of batches of pipelines to search. Parameters `max_time`, and `max_iterations` have precedence over stopping the search.
- **problem_configuration** (*dict*, *None*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, `forecast_horizon`, and `max_delay` variables.
- **train_best_pipeline** (*boolean*) – Whether or not to train the best pipeline before returning it. Defaults to True.
- **pipeline_parameters** (*dict*) – A dict of the parameters used to initialize a pipeline with. Keys should consist of the component names and values should specify parameter values
e.g. `pipeline_parameters = { 'Imputer' : { 'numeric_impute_strategy': 'most_frequent' } }`
- **custom_hyperparameters** (*dict*) – A dict of the hyperparameter ranges used to iterate over during search. Keys should consist of the component names and values should specify a singular value or `skopt.Space`.
e.g. `custom_hyperparameters = { 'Imputer' : { 'numeric_impute_strategy': Categorical(['most_frequent', 'median']) } }`
- **sampler_method** (*str*) – The data sampling component to use in the pipelines if the problem type is classification and the target balance is smaller than the `sampler_balanced_ratio`. Either 'auto', which will use our preferred sampler for the data, 'Undersampler', 'Oversampler', or None. Defaults to 'auto'.
- **sampler_balanced_ratio** (*float*) – The minority:majority class ratio that we consider balanced, so a 1:4 ratio would be equal to 0.25. If the class balance is larger than this provided value, then we will not add a sampler since the data is then considered balanced. Overrides the `sampler_ratio` of the samplers. Defaults to 0.25.
- **_ensembling_split_size** (*float*) – The amount of the training data we'll set aside for training ensemble metalearners. Only used when ensembling is True. Must be between 0 and 1, exclusive. Defaults to 0.2
- **_pipelines_per_batch** (*int*) – The number of pipelines to train for every batch after the first one. The first batch will train a baseline pipeline + one of each pipeline family allowed in the search.
- **_automl_algorithm** (*str*) – The automl algorithm to use. Currently the two choices are 'iterative' and 'default'. Defaults to *iterative*.
- **engine** (*EngineBase* or *str*) – The engine instance used to evaluate pipelines. Dask or `concurrent.futures` engines can also be chosen by providing a string from the list ["sequential", "cf_threaded", "cf_process", "dask_threaded", "dask_process"]. If a parallel engine is selected this way, the maximum amount of parallelism, as determined by the engine, will be used. Defaults to "sequential".
- **verbose** (*boolean*) – Whether or not to display semi-real-time updates to stdout while search is running. Defaults to False.

Methods

<code>add_to_rankings</code>	Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.
<code>best_pipeline</code>	Returns a trained instance of the best pipeline and parameters found during automl search. If <code>train_best_pipeline</code> is set to False, returns an untrained pipeline instance.
<code>close_engine</code>	Function to explicitly close the engine, client, parallel resources.
<code>describe_pipeline</code>	Describe a pipeline
<code>full_rankings</code>	Returns a pandas.DataFrame with scoring results from all pipelines searched
<code>get_pipeline</code>	Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline
<code>load</code>	Loads AutoML object at file path
<code>plot</code>	
<code>rankings</code>	Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.
<code>results</code>	Class that allows access to a copy of the results from <code>automl_search</code> .
<code>save</code>	Saves AutoML object at file path
<code>score_pipelines</code>	Score a list of pipelines on the given holdout data.
<code>search</code>	Find the best pipeline for the data set.
<code>train_pipelines</code>	Train a list of pipelines on the training data.

add_to_rankings (*self*, *pipeline*)

Fits and evaluates a given pipeline then adds the results to the automl rankings with the requirement that automl search has been run.

Parameters **pipeline** (*PipelineBase*) – pipeline to train and evaluate.

property best_pipeline (*self*)

Returns a trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Returns A trained instance of the best pipeline and parameters found during automl search. If `train_best_pipeline` is set to False, returns an untrained pipeline instance.

Return type *PipelineBase*

close_engine (*self*)

Function to explicitly close the engine, client, parallel resources.

describe_pipeline (*self*, *pipeline_id*, *return_dict=False*)

Describe a pipeline

Parameters

- **pipeline_id** (*int*) – pipeline to describe
- **return_dict** (*bool*) – If True, return dictionary of information about pipeline. Defaults to False.

Returns Description of specified pipeline. Includes information such as type of pipeline components, problem, training time, cross validation, etc.

property full_rankings (*self*)

Returns a pandas.DataFrame with scoring results from all pipelines searched

get_pipeline (*self*, *pipeline_id*)

Given the ID of a pipeline training result, returns an untrained instance of the specified pipeline initialized with the parameters used to train that pipeline during automl search.

Parameters *pipeline_id* (*int*) – pipeline to retrieve

Returns untrained pipeline instance associated with the provided ID

Return type PipelineBase

static load (*file_path*, *pickle_type*='cloudpickle')

Loads AutoML object at file path

Parameters

- **file_path** (*str*) – location to find file to load
- {"pickle" (*pickle_type*) – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.
- "cloudpickle"} – the pickling library to use. Currently not used since the standard pickle library can handle cloudpickles.

Returns AutoSearchBase object

property plot (*self*)

property rankings (*self*)

Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.

property results (*self*)

Class that allows access to a copy of the results from *automl_search*.

Returns: dict containing *pipeline_results*: a dict with results from each pipeline, and *search_order*: a list describing the order the pipelines were searched.

save (*self*, *file_path*, *pickle_type*='cloudpickle', *pickle_protocol*=cloudpickle.DEFAULT_PROTOCOL)

Saves AutoML object at file path

Parameters

- **file_path** (*str*) – location to save file
- {"pickle" (*pickle_type*) – the pickling library to use.
- "cloudpickle"} – the pickling library to use.
- **pickle_protocol** (*int*) – the pickle data stream format.

Returns None

score_pipelines (*self*, *pipelines*, *X_holdout*, *y_holdout*, *objectives*)

Score a list of pipelines on the given holdout data.

Parameters

- **pipelines** (*list* (*PipelineBase*)) – List of pipelines to train.
- **X_holdout** (*pd.DataFrame*) – Holdout features.
- **y_holdout** (*pd.Series*) – Holdout targets for scoring.

- **objectives** (*list(str), list(ObjectiveBase)*) – Objectives used for scoring.

Returns Dictionary keyed by pipeline name that maps to a dictionary of scores. Note that the any pipelines that error out during scoring will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, Dict[str, float]]

search (*self, show_iteration_plot=True*)

Find the best pipeline for the data set.

Parameters

- **feature_types** (*list, optional*) – list of feature types, either numerical or categorical. Categorical features will automatically be encoded
- **show_iteration_plot** (*boolean, True*) – Shows an iteration vs. score plot in Jupyter notebook. Disabled by default in non-Jupyter environments.

train_pipelines (*self, pipelines*)

Train a list of pipelines on the training data.

This can be helpful for training pipelines once the search is complete.

Parameters **pipelines** (*list(PipelineBase)*) – List of pipelines to train.

Returns Dictionary keyed by pipeline name that maps to the fitted pipeline. Note that the any pipelines that error out during training will not be included in the dictionary but the exception and stacktrace will be displayed in the log.

Return type Dict[str, PipelineBase]

`evalml.search(X_train=None, y_train=None, problem_type=None, objective='auto', mode='fast', max_time=None, patience=None, tolerance=None, problem_configuration=None)`

Given data and configuration, run an automl search.

This method will run EvalML's default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you'd like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str or ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.
- **objective** (*str, ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to 'auto', chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.

- **mode** (*str*) – mode for DefaultAlgorithm. There are two modes: fast and long, where fast is a subset of long. Please look at DefaultAlgorithm for more details.
- **max_time** (*int*, *str*) – Maximum time to search for pipelines. This will not start a new pipeline search after the duration has elapsed. If it is an integer, then the time will be in seconds. For strings, time can be specified as seconds, minutes, or hours.
- **patience** (*int*) – Number of iterations without improvement to stop search early. Must be positive. If None, early stopping is disabled. Defaults to None.
- **tolerance** (*float*) – Minimum percentage difference to qualify as score improvement for early stopping. Only applicable if patience is not None. Defaults to None.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example, in time series problems, values should be passed in for the `date_index`, `gap`, and `max_delay` variables.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

```
evalml.search_iterative(X_train=None, y_train=None, problem_type=None, objective='auto',
                        problem_configuration=None, **kwargs)
```

Given data and configuration, run an automl search.

This method will run EvalML's default suite of data checks. If the data checks produce errors, the data check results will be returned before running the automl search. In that case we recommend you alter your data to address these errors and try again.

This method is provided for convenience. If you'd like more control over when each of these steps is run, consider making calls directly to the various pieces like the data checks and AutoMLSearch, instead of using this method.

Parameters

- **X_train** (*pd.DataFrame*) – The input training data of shape [n_samples, n_features]. Required.
- **y_train** (*pd.Series*) – The target training data of length [n_samples]. Required for supervised learning tasks.
- **problem_type** (*str* or *ProblemTypes*) – Type of supervised learning problem. See `evalml.problem_types.ProblemType.all_problem_types` for a full list.
- **objective** (*str*, *ObjectiveBase*) – The objective to optimize for. Used to propose and rank pipelines, but not for optimizing each pipeline during fit-time. When set to 'auto', chooses:
 - LogLossBinary for binary classification problems,
 - LogLossMulticlass for multiclass classification problems, and
 - R2 for regression problems.
- **problem_configuration** (*dict*) – Additional parameters needed to configure the search. For example,
- **time series problems** (*in*) –
- **should be passed in for the date_index** (*values*) –
- **gap** –

- `forecast_horizon` –
- `max_delay variables.` (*and*) –

Other keyword arguments which are provided will be passed to `AutoMLSearch`.

Returns the automl search object containing pipelines and rankings, and the results from running the data checks. If the data check results contain errors, automl search will not be run and an automl search object will not be returned.

Return type (*AutoMLSearch*, dict)

RELEASE NOTES

Future Releases

- Enhancements
- Fixes
- Changes
- Testing Changes

Warning: Breaking Changes

v0.32.1 Sep. 10, 2021

- **Enhancements**
 - Added `verbose` flag to `AutoMLSearch` to run search in silent mode by default [#2645](#)
 - Added label encoder to `XGBoostClassifier` to remove the warning [#2701](#)
 - Set `eval_metric` to `logloss` for `XGBoostClassifier` [#2741](#)
 - Added support for `woodwork` versions 0.7.0 and 0.7.1 [#2743](#)
 - Changed `explain_predictions` functions to display original feature values [#2759](#)
 - Added `X_train` and `y_train` to `graph_prediction_vs_actual_over_time` and `get_prediction_vs_actual_over_time_data` [#2762](#)
 - Added `forecast_horizon` as a required parameter to time series pipelines and `AutoMLSearch` [#2697](#)
 - Added `predict_in_sample` and `predict_proba_in_sample` methods to time series pipelines to predict on data where the target is known, e.g. cross-validation [#2697](#)
- **Fixes**
 - Fixed bug where `_catch_warnings` assumed all warnings were `PipelineNotUsed` [#2753](#)
 - Fixed bug where `Imputer.transform` would erase `ww` typing information prior to handing data to the `SimpleImputer` [#2752](#)
 - Fixed bug where `Oversampler` could not be copied [#2755](#)
- **Changes**
 - Deleted `drop_nan_target_rows` utility method [#2737](#)
 - Removed default logging setup and debugging log file [#2645](#)

- Changed the default `n_jobs` value for `XGBoostClassifier` and `XGBoostRegressor` to 12 [#2757](#)
- Changed `TimeSeriesBaselineEstimator` to only work on a time series pipeline with a `DelayedFeaturesTransformer` [#2697](#)
- Added `X_train` and `y_train` as optional parameters to pipeline `predict`, `predict_proba`. Only used for time series pipelines [#2697](#)
- Added `training_data` and `training_target` as optional parameters to `explain_predictions` and `explain_predictions_best_worst` to support time series pipelines [#2697](#)
- Changed time series pipeline predictions to no longer output series/dataframes padded with NaNs. A prediction will be returned for every row in the `X` input [#2697](#)
- **Documentation Changes**
 - Specified installation steps for Prophet [#2713](#)
 - Added documentation for data exploration on data check actions [#2696](#)
 - Added a user guide entry for time series modelling [#2697](#)
- **Testing Changes**
 - Fixed flaky `TargetDistributionDataCheck` test for very_lognormal distribution [#2748](#)

Warning:

Breaking Changes

- Removed default logging setup and debugging log file [#2645](#)
- Added `X_train` and `y_train` to `graph_prediction_vs_actual_over_time` and `get_prediction_vs_actual_over_time_data` [#2762](#)
- Added `forecast_horizon` as a required parameter to time series pipelines and `AutoMLSearch` [#2697](#)
- Changed `TimeSeriesBaselineEstimator` to only work on a time series pipeline with a `DelayedFeaturesTransformer` [#2697](#)
- Added `X_train` and `y_train` as required parameters for `predict` and `predict_proba` in time series pipelines [#2697](#)
- Added `training_data` and `training_target` as required parameters to `explain_predictions` and `explain_predictions_best_worst` for time series pipelines [#2697](#)

v0.32.0 Aug. 31, 2021

- **Enhancements**
 - Allow string for engine parameter for `AutoMLSearch` [#2667](#)
 - Add `ProphetRegressor` to `AutoML` [#2619](#)
 - Integrated `DefaultAlgorithm` into `AutoMLSearch` [#2634](#)
 - Removed SVM “linear” and “precomputed” kernel hyperparameter options, and improved default parameters [#2651](#)

- Updated `ComponentGraph` initialization to raise `ValueError` when user attempts to use `.y` for a component that does not produce a tuple output [#2662](#)
- Updated to support Woodwork 0.6.0 [#2690](#)
- Updated pipeline `graph()` to distinguish X and y edges [#2654](#)
- Added `DropRowsTransformer` component [#2692](#)
- Added `DROP_ROWS` to `_make_component_list_from_actions` and clean up metadata [#2694](#)
- **Fixes**
 - Updated `Oversampler` logic to select best SMOTE based on component input instead of pipeline input [#2695](#)
 - Added ability to explicitly close `DaskEngine` resources to improve runtime and reduce Dask warnings [#2667](#)
 - Fixed partial dependence bug for ensemble pipelines [#2714](#)
 - Updated `TargetLeakageDataCheck` to maintain user-selected logical types [#2711](#)
- **Changes**
 - Replaced `SMOTEOverSampler`, `SMOTENOverSampler` and `SMOTENCOversampler` with consolidated `Oversampler` component [#2695](#)
 - Removed `LinearRegressor` from the list of default `AutoMLSearch` estimators due to poor performance [#2660](#)
- **Documentation Changes**
 - Added user guide documentation for using `ComponentGraph` and added `ComponentGraph` to API reference [#2673](#)
 - Updated documentation to make parallelization of AutoML clearer [#2667](#)
- **Testing Changes**
 - Removes the process-level parallelism from the `test_cancel_job` test [#2666](#)
 - Installed numba 0.53 in windows CI to prevent problems installing version 0.54 [#2710](#)

Warning:**Breaking Changes**

- Renamed the current top level search method to `search_iterative` and defined a new search method for the `DefaultAlgorithm` [#2634](#)
- Replaced `SMOTEOverSampler`, `SMOTENOverSampler` and `SMOTENCOversampler` with consolidated `Oversampler` component [#2695](#)
- Removed `LinearRegressor` from the list of default `AutoMLSearch` estimators due to poor performance [#2660](#)

v0.31.0 Aug. 19, 2021

- **Enhancements**
 - Updated the high variance check in `AutoMLSearch` to be robust to a variety of objectives and cv scores [#2622](#)
 - Use Woodwork's outlier detection for the `OutliersDataCheck` [#2637](#)

- Added ability to utilize instantiated components when creating a pipeline #2643
- Sped up the all Nan and unknown check in `infer_feature_types` #2661
- Fixes
- Changes
 - Deleted `_put_into_original_order` helper function #2639
 - Refactored time series pipeline code using a time series pipeline base class #2649
 - Renamed `dask_tests` to `parallel_tests` #2657
 - Removed commented out code in `pipeline_meta.py` #2659
- Documentation Changes
 - Add complete install command to README and Install section #2627
 - Cleaned up documentation for `MulticollinearityDataCheck` #2664
- Testing Changes
 - Speed up CI by splitting Prophet tests into a separate workflow in GitHub #2644

Warning:

Breaking Changes

- `TimeSeriesRegressionPipeline` no longer inherits from `TimeSeriesRegressionPipeline` #2649

v0.30.2 Aug. 16, 2021

- Fixes
 - Updated changelog and version numbers to match the release. Release 0.30.1 was release erroneously without a change to the version numbers. 0.30.2 replaces it.

v0.30.1 Aug. 12, 2021

- Enhancements
 - Added `DatetimeFormatDataCheck` for time series problems #2603
 - Added `ProphetRegressor` to estimators #2242
 - Updated `ComponentGraph` to handle not calling samplers' transform during predict, and updated samplers' transform methods s.t. `fit_transform` is equivalent to `fit(X, y).transform(X, y)` #2583
 - Updated `ComponentGraph_validate_component_dict` logic to be stricter about input values #2599
 - Patched bug in `xgboost` estimators where predicting on a feature matrix of only booleans would throw an exception. #2602
 - Updated `ARIMAREgressor` to use relative forecasting to predict values #2613
 - Added support for creating pipelines without an estimator as the final component and added `transform(X, y)` method to pipelines and component graphs #2625
 - Updated to support Woodwork 0.5.1 #2610
- Fixes

- Updated `AutoMLSearch` to drop `ARIMAREgressor` from `allowed_estimators` if an incompatible frequency is detected [#2632](#)
- Updated `get_best_sampler_for_data` to consider all non-numeric datatypes as categorical for SMOTE [#2590](#)
- Fixed inconsistent test results from `TargetDistributionDataCheck` [#2608](#)
- Adopted vectorized `pd.NA` checking for Woodwork 0.5.1 support [#2626](#)
- Pinned upper version of `astroid` to 2.6.6 to keep `ReadTheDocs` working. [#2638](#)

- **Changes**

- Renamed SMOTE samplers to SMOTE oversampler [#2595](#)
- Changed `partial_dependence` and `graph_partial_dependence` to raise a `PartialDependenceError` instead of `ValueError`. This is not a breaking change because `PartialDependenceError` is a subclass of `ValueError` [#2604](#)
- Cleaned up code duplication in `ComponentGraph` [#2612](#)
- Stored `predict_proba` results in `.x` for intermediate estimators in `ComponentGraph` [#2629](#)

- **Documentation Changes**

- To avoid local docs build error, only add warning disable and download headers on `ReadTheDocs` builds, not locally [#2617](#)

- **Testing Changes**

- Updated `partial_dependence` tests to change the element-wise comparison per the Plotly 5.2.1 upgrade [#2638](#)
- Changed the lint CI job to only check against python 3.9 via the `-t` flag [#2586](#)
- Installed Prophet in linux nightlies test and fixed `test_all_components` [#2598](#)
- Refactored and fixed all `make_pipeline` tests to assert correct order and address new Woodwork Unknown type inference [#2572](#)
- Removed `component_graphs` as a global variable in `test_component_graphs.py` [#2609](#)

Warning:
Breaking Changes

- Renamed SMOTE samplers to SMOTE oversampler. Please use `SMOTEOverSampler`, `SMOTENCOversampler`, `SMOTENOverSampler` instead of `SMOTESampler`, `SMOTENCSampler`, and `SMOTENSampler` [#2595](#)

v0.30.0 Aug. 3, 2021

- **Enhancements**

- Added `LogTransformer` and `TargetDistributionDataCheck` [#2487](#)
- Issue a warning to users when a pipeline parameter passed in isn't used in the pipeline [#2564](#)
- Added Gini coefficient as an objective [#2544](#)
- Added `repr` to `ComponentGraph` [#2565](#)
- Added components to extract features from URL and EmailAddress Logical Types [#2550](#)

- Added support for *NaN* values in `TextFeaturizer` #2532
- Added `SelectByType` transformer #2531
- Added separate thresholds for percent null rows and columns in `HighlyNullDataCheck` #2562
- Added support for *NaN* natural language values #2577
- **Fixes**
 - Raised error message for types `URL`, `NaturalLanguage`, and `EmailAddress` in `partial_dependence` #2573
- **Changes**
 - Updated `PipelineBase` implementation for creating pipelines from a list of components #2549
 - Moved `get_hyperparameter_ranges` to `PipelineBase` class from `automl/utlis` module #2546
 - Renamed `ComponentGraph`’s `get_parents` to `get_inputs` #2540
 - Removed `ComponentGraph.linearized_component_graph` and `ComponentGraph.from_list` #2556
 - Updated `ComponentGraph` to enforce requiring `.x` and `.y` inputs for each component in the graph #2563
 - Renamed existing ensembler implementation from `StackedEnsemblers` to `SklearnStackedEnsemblers` #2578
- **Documentation Changes**
 - Added documentation for `DaskEngine` and `CFEngine` parallel engines #2560
 - Improved detail of `TextFeaturizer` docstring and tutorial #2568
- **Testing Changes**
 - Added test that makes sure `split_data` does not shuffle for time series problems #2552

Warning:

Breaking Changes

- Moved `get_hyperparameter_ranges` to `PipelineBase` class from `automl/utlis` module #2546
- Renamed `ComponentGraph`’s `get_parents` to `get_inputs` #2540
- Removed `ComponentGraph.linearized_component_graph` and `ComponentGraph.from_list` #2556
- Updated `ComponentGraph` to enforce requiring `.x` and `.y` inputs for each component in the graph #2563

v0.29.0 Jul. 21, 2021

- **Enhancements**
 - Updated 1-way partial dependence support for datetime features #2454
 - Added details on how to fix error caused by broken `ww` schema #2466

- Added ability to use built-in pickle for saving AutoMLSearch #2463
- Updated our components and component graphs to use latest features of ww 0.4.1, e.g. `concat_columns` and `drop in-place`. #2465
- Added new, concurrent.futures based engine for parallel AutoML #2506
- Added support for new Woodwork Unknown type in AutoMLSearch #2477
- Updated our components with an attribute that describes if they modify features or targets and can be used in list API for pipeline initialization #2504
- Updated ComponentGraph to accept X and y as inputs #2507
- Removed unused TARGET_BINARY_INVALID_VALUES from DataCheckMessageCode enum and fixed formatting of objective documentation #2520
- Added EvalMLAlgorithm #2525
- Added support for *NaN* values in TextFeaturizer #2532

- **Fixes**

- Fixed FraudCost objective and reverted threshold optimization method for binary classification to Golden #2450
- Added custom exception message for partial dependence on features with scales that are too small #2455
- Ensures the typing for Ordinal and Datetime ltypes are passed through `_retain_custom_types_and_initialize_woodwork` #2461
- Updated to work with Pandas 1.3.0 #2442
- Updated to work with sktime 0.7.0 #2499

- **Changes**

- Updated XGBoost dependency to `>=1.4.2` #2484, #2498
- Added a DeprecationWarning about deprecating the list API for ComponentGraph #2488
- Updated `make_pipeline` for AutoML to create dictionaries, not lists, to initialize pipelines #2504
- No longer installing graphviz on windows in our CI pipelines because release 0.17 breaks windows 3.7 #2516

- **Documentation Changes**

- Moved docstrings from `__init__` to class pages, added missing docstrings for missing classes, and updated missing default values #2452
- Build documentation with sphinx-autoapi #2458
- Change `autoapi_ignore` to only ignore files in `evalml/tests/*` #2530

- **Testing Changes**

- Fixed flaky dask tests #2471
- Removed shellcheck action from `build_conda_pkg` action #2514
- Added a `tmp_dir` fixture that deletes its contents after tests run #2505
- Added a test that makes sure all pipelines in AutoMLSearch get the same data splits #2513

- Condensed warning output in test logs #2521

Warning:

Breaking Changes

- *NaN* values in the *Natural Language* type are no longer supported by the Imputer with the pandas upgrade. #2477

v0.28.0 Jul. 2, 2021

- **Enhancements**

- Added support for showing a Individual Conditional Expectations plot when graphing Partial Dependence #2386
- Exposed `thread_count` for Catboost estimators as `n_jobs` parameter #2410
- Updated Objectives API to allow for sample weighting #2433

- **Fixes**

- Deleted unreachable line from `IterativeAlgorithm` #2464

- **Changes**

- Pinned Woodwork version between 0.4.1 and 0.4.2 #2460
- Updated psutils minimum version in requirements #2438
- Updated `log_error_callback` to not include filepath in logged message #2429

- **Documentation Changes**

- Sped up docs #2430
- Removed mentions of `DataTable` and `DataColumn` from the docs #2445

- **Testing Changes**

- Added slack integration for nightlies tests #2436
- Changed `build_conda_pkg` CI job to run only when dependencies are updates #2446
- Updated workflows to store pytest runtimes as test artifacts #2448
- Added `AutoMLTestEnv` test fixture for making it easy to mock automl tests #2406

v0.27.0 Jun. 22, 2021

- **Enhancements**

- Adds force plots for prediction explanations #2157
- Removed self-reference from `AutoMLSearch` #2304
- Added support for nonlinear pipelines for `generate_pipeline_code` #2332
- Added `inverse_transform` method to pipelines #2256
- Add optional automatic update checker #2350
- Added `search_order` to `AutoMLSearch`'s rankings and `full_rankings` tables #2345
- Updated threshold optimization method for binary classification #2315
- Updated demos to pull data from S3 instead of including demo data in package #2387

- Upgrade woodwork version to v0.4.1 [#2379](#)
- **Fixes**
 - Preserve user-specified woodwork types throughout pipeline fit/predict [#2297](#)
 - Fixed ComponentGraph appending target to final_component_features if there is a component that returns both X and y [#2358](#)
 - Fixed partial dependence graph method failing on multiclass problems when the class labels are numeric [#2372](#)
 - Added thresholding_objective argument to AutoMLSearch for binary classification problems [#2320](#)
 - Added change for k_neighbors parameter in SMOTE Oversamplers to automatically handle small samples [#2375](#)
 - Changed naming for Logistic Regression Classifier file [#2399](#)
 - Pinned pytest-timeout to fix minimum dependence checker [#2425](#)
 - Replaced Elastic Net Classifier base class with Logistic Regression to avoid NaN outputs [#2420](#)
- **Changes**
 - Cleaned up PipelineBase's component_graph and _component_graph attributes. Updated PipelineBase __repr__ and added __eq__ for ComponentGraph [#2332](#)
 - Added and applied black linting package to the EvalML repo in place of autopep8 [#2306](#)
 - Separated custom_hyperparameters from pipelines and added them as an argument to AutoMLSearch [#2317](#)
 - Replaced allowed_pipelines with allowed_component_graphs [#2364](#)
 - Removed private method _compute_features_during_fit from PipelineBase [#2359](#)
 - Updated compute_order in ComponentGraph to be a read-only property [#2408](#)
 - Unpinned PyZMQ version in requirements.txt [#2389](#)
 - Uncapping LightGBM version in requirements.txt [#2405](#)
 - Updated minimum version of plotly [#2415](#)
 - Removed SensitivityLowAlert objective from core objectives [#2418](#)
- **Documentation Changes**
 - Fixed lead scoring weights in the demos documentation [#2315](#)
 - Fixed start page code and description dataset naming discrepancy [#2370](#)
- **Testing Changes**
 - Update minimum unit tests to run on all pull requests [#2314](#)
 - Pass token to authorize uploading of codecov reports [#2344](#)
 - Add pytest-timeout. All tests that run longer than 6 minutes will fail. [#2374](#)
 - Separated the dask tests out into separate github action jobs to isolate dask failures. [#2376](#)
 - Refactored dask tests [#2377](#)
 - Added the combined dask/non-dask unit tests back and renamed the dask only unit tests. [#2382](#)

- Sped up unit tests and split into separate jobs [#2365](#)
- Change CI job names, run lint for python 3.9, run nightlies on python 3.8 at 3am EST [#2395](#) [#2398](#)
- Set fail-fast to false for CI jobs that run for PRs [#2402](#)

Warning:

Breaking Changes

- *AutoMLSearch* will accept *allowed_component_graphs* instead of *allowed_pipelines* [#2364](#)
- Removed *PipelineBase*'s *_component_graph* attribute. Updated *PipelineBase* *__repr__* and added *__eq__* for *ComponentGraph* [#2332](#)
- *pipeline_parameters* will no longer accept *skopt.space* variables since hyperparameter ranges will now be specified through *custom_hyperparameters* [#2317](#)

v0.25.0 Jun. 01, 2021

• **Enhancements**

- Upgraded minimum woodwork to version 0.3.1. Previous versions will not be supported [#2181](#)
- Added a new callback parameter for *explain_predictions_best_worst* [#2308](#)

• **Fixes**

• **Changes**

- Deleted the *return_pandas* flag from our demo data loaders [#2181](#)
- Moved *default_parameters* to *ComponentGraph* from *PipelineBase* [#2307](#)

• **Documentation Changes**

- Updated the release procedure documentation [#2230](#)

• **Testing Changes**

- Ignoring *test_saving_png_file* while building conda package [#2323](#)

Warning:

Breaking Changes

- Deleted the *return_pandas* flag from our demo data loaders [#2181](#)
- Upgraded minimum woodwork to version 0.3.1. Previous versions will not be supported [#2181](#)
- Due to the weak-ref in woodwork, set the result of *infer_feature_types* to a variable before accessing woodwork [#2181](#)

v0.24.2 May. 24, 2021

• **Enhancements**

- Added oversamplers to *AutoMLSearch* [#2213](#) [#2286](#)
- Added dictionary input functionality for *Undersampler* component [#2271](#)
- Changed the default parameter values for *Elastic Net Classifier* and *Elastic Net Regressor* [#2269](#)

- Added dictionary input functionality for the Oversampler components [#2288](#)
- **Fixes**
 - Set default `n_jobs` to 1 for `StackedEnsembleClassifier` and `StackedEnsembleRegressor` until fix for text-based parallelism in sklearn stacking can be found [#2295](#)
- **Changes**
 - Updated `start_iteration_callback` to accept a pipeline instance instead of a pipeline class and no longer accept pipeline parameters as a parameter [#2290](#)
 - Refactored `calculate_permutation_importance` method and add per-column permutation importance method [#2302](#)
 - Updated logging information in `AutoMLSearch.__init__` to clarify pipeline generation [#2263](#)
- **Documentation Changes**
 - Minor changes to the release procedure [#2230](#)
- **Testing Changes**
 - Use codecov action to update coverage reports [#2238](#)
 - Removed MarkupSafe dependency version pin from requirements.txt and moved instead into RTD docs build CI [#2261](#)

Warning:**Breaking Changes**

- Updated `start_iteration_callback` to accept a pipeline instance instead of a pipeline class and no longer accept pipeline parameters as a parameter [#2290](#)
- Moved `default_parameters` to `ComponentGraph` from `PipelineBase`. A pipeline's `default_parameters` is now accessible via `pipeline.component_graph.default_parameters` [#2307](#)

v0.24.1 May. 16, 2021

- **Enhancements**
 - Integrated `ARIMAREgressor` into `AutoML` [#2009](#)
 - Updated `HighlyNullDataCheck` to also perform a null row check [#2222](#)
 - Set `max_depth` to 1 in calls to `featuretools dfs` [#2231](#)
- **Fixes**
 - Removed data splitter sampler calls during training [#2253](#)
 - Set minimum required version for `pyzmq`, `colorama`, and `docutils` [#2254](#)
 - Changed `BaseSampler` to return `None` instead of `y` [#2272](#)
- **Changes**
 - Removed ensemble split and indices in `AutoMLSearch` [#2260](#)
 - Updated pipeline `repr()` and `generate_pipeline_code` to return pipeline instances without generating custom pipeline class [#2227](#)
- **Documentation Changes**

- Capped Sphinx version under 4.0.0 #2244
- **Testing Changes**
 - Change number of cores for pytest from 4 to 2 #2266
 - Add minimum dependency checker to generate minimum requirement files #2267
 - Add unit tests with minimum dependencies #2277

v0.24.0 May. 04, 2021

- **Enhancements**
 - Added `date_index` as a required parameter for TimeSeries problems #2217
 - Have the `OneHotEncoder` return the transformed columns as booleans rather than floats #2170
 - Added Oversampler transformer component to EvalML #2079
 - Added Undersampler to `AutoMLSearch`, as well as arguments `_sampler_method` and `sampler_balanced_ratio` #2128
 - Updated prediction explanations functions to allow pipelines with XGBoost estimators #2162
 - Added partial dependence for datetime columns #2180
 - Update precision-recall curve with positive label index argument, and fix for 2d predicted probabilities #2090
 - Add `pct_null_rows` to `HighlyNullDataCheck` #2211
 - Added a standalone `AutoML search` method for convenience, which runs data checks and then runs `automl` #2152
 - Make the first batch of AutoML have a predefined order, with linear models first and complex models last #2223 #2225
 - Added sampling dictionary support to `BalancedClassificationSampler` #2235
- **Fixes**
 - Fixed partial dependence not respecting grid resolution parameter for numerical features #2180
 - Enable prediction explanations for catboost for multiclass problems #2224
- **Changes**
 - Deleted baseline pipeline classes #2202
 - Reverting user specified date feature PR #2155 until `pmdarima` installation fix is found #2214
 - Updated pipeline API to accept component graph and other class attributes as instance parameters. Old pipeline API still works but will not be supported long-term. #2091
 - Removed all old datasplitters from EvalML #2193
 - Deleted `make_pipeline_from_components` #2218
- **Documentation Changes**
 - Renamed dataset to clarify that its gzipped but not a tarball #2183
 - Updated documentation to use pipeline instances instead of pipeline subclasses #2195
 - Updated contributing guide with a note about GitHub Actions permissions #2090
 - Updated `automl` and model understanding user guides #2090
- **Testing Changes**

- Use machineFL user token for dependency update bot, and add more reviewers [#2189](#)

Warning:**Breaking Changes**

- All `baseline` `pipeline` `classes` (`BaselineBinaryPipeline`, `BaselineMulticlassPipeline`, `BaselineRegressionPipeline`, etc.) have been deleted [#2202](#)
- Updated pipeline API to accept component graph and other class attributes as instance parameters. Old pipeline API still works but will not be supported long-term. Pipelines can now be initialized by specifying the component graph as the first parameter, and then passing in optional arguments such as `custom_name`, `parameters`, etc. For example, `BinaryClassificationPipeline(["Random Forest Classifier"], parameters={})`. [#2091](#)
- Removed all old datasplitters from EvalML [#2193](#)
- Deleted utility method `make_pipeline_from_components` [#2218](#)

v0.23.0 Apr. 20, 2021• **Enhancements**

- Refactored `EngineBase` and `SequentialEngine` api. Adding `DaskEngine` [#1975](#).
- Added optional engine argument to `AutoMLSearch` [#1975](#)
- Added a warning about how time series support is still in beta when a user passes in a time series problem to `AutoMLSearch` [#2118](#)
- Added `NaturalLanguageNaNDataCheck` data check [#2122](#)
- Added `ValueError` to `partial_dependence` to prevent users from computing partial dependence on columns with all NaNs [#2120](#)
- Added standard deviation of cv scores to rankings table [#2154](#)

• **Fixes**

- Fixed `BalancedClassificationDataCVSplit`, `BalancedClassificationDataTVSplit`, and `BalancedClassificationSampler` to use `minority:majority` ratio instead of `majority:minority` [#2077](#)
- Fixed bug where two-way partial dependence plots with categorical variables were not working correctly [#2117](#)
- Fixed bug where hyperparameters were not displaying properly for pipelines with a list `component_graph` and duplicate components [#2133](#)
- Fixed bug where `pipeline_parameters` argument in `AutoMLSearch` was not applied to pipelines passed in as `allowed_pipelines` [#2133](#)
- Fixed bug where `AutoMLSearch` was not applying custom hyperparameters to pipelines with a list `component_graph` and duplicate components [#2133](#)

• **Changes**

- Removed `hyperparameter_ranges` from `Undersampler` and renamed `balanced_ratio` to `sampling_ratio` for samplers [#2113](#)
- Renamed `TARGET_BINARY_NOT_TWO_EXAMPLES_PER_CLASS` data check message code to `TARGET_MULTICLASS_NOT_TWO_EXAMPLES_PER_CLASS` [#2126](#)

- Modified one-way partial dependence plots of categorical features to display data with a bar plot #2117
- Renamed `score` column for `automl.rankings` as `mean_cv_score` #2135
- Remove ‘warning’ from docs tool output #2031

- **Documentation Changes**

- Fixed `conf.py` file #2112
- Added a sentence to the automl user guide stating that our support for time series problems is still in beta. #2118
- Fixed documentation demos #2139
- Update test badge in README to use GitHub Actions #2150

- **Testing Changes**

- Fixed `test_describe_pipeline` for pandas v1.2.4 #2129
- Added a GitHub Action for building the conda package #1870 #2148

Warning:

Breaking Changes

- Renamed `balanced_ratio` to `sampling_ratio` for the `BalancedClassificationDataCVSplit`, `BalancedClassificationDataTVSplit`, `BalancedClassificationSampler`, and `Undersampler` #2113
- Deleted the “errors” key from automl results #1975
- Deleted the `raise_and_save_error_callback` and the `log_and_save_error_callback` #1975
- Fixed `BalancedClassificationDataCVSplit`, `BalancedClassificationDataTVSplit`, and `BalancedClassificationSampler` to use `minority:majority` ratio instead of `majority:minority` #2077

v0.22.0 Apr. 06, 2021

- **Enhancements**

- Added a GitHub Action for `linux_unit_tests` #2013
- Added recommended actions for `InvalidTargetDataCheck`, updated `_make_component_list_from_actions` to address new action, and added `TargetImputer` component #1989
- Updated `AutoMLSearch._check_for_high_variance` to not emit `RuntimeWarning` #2024
- Added exception when pipeline passed to `explain_predictions` is a `Stacked Ensemble` pipeline #2033
- Added sensitivity at low alert rates as an objective #2001
- Added `Undersampler` transformer component #2030

- **Fixes**

- Updated Engine’s `train_batch` to apply undersampling #2038

- Fixed bug in where Time Series Classification pipelines were not encoding targets in `predict` and `predict_proba` #2040
- Fixed data splitting errors if target is float for classification problems #2050
- Pinned `docutils` to `<0.17` to fix `ReadtheDocs` warning issues #2088

- **Changes**

- Removed lists as acceptable hyperparameter ranges in `AutoMLSearch` #2028
- Renamed “details” to “metadata” for data check actions #2008

- **Documentation Changes**

- Catch and suppress warnings in documentation #1991 #2097
- Change spacing in `start.ipynb` to provide clarity for `AutoMLSearch` #2078
- Fixed start code on README #2108

- Testing Changes

v0.21.0 Mar. 24, 2021

- **Enhancements**

- Changed `AutoMLSearch` to default `optimize_thresholds` to `True` #1943
- Added multiple oversampling and undersampling sampling methods as data splitters for imbalanced classification #1775
- Added params to balanced classification data splitters for visibility #1966
- Updated `make_pipeline` to not add `Imputer` if input data does not have numeric or categorical columns #1967
- Updated `ClassImbalanceDataCheck` to better handle multiclass imbalances #1986
- Added recommended actions for the output of data check’s `validate` method #1968
- Added error message for `partial_dependence` when features are mostly the same value #1994
- Updated `OneHotEncoder` to drop one redundant feature by default for features with two categories #1997
- Added a `PolynomialDetrender` component #1992
- Added `DateTimeNaNDataCheck` data check #2039

- **Fixes**

- Changed best pipeline to train on the entire dataset rather than just ensemble indices for ensemble problems #2037
- Updated binary classification pipelines to use objective decision function during scoring of custom objectives #1934

- **Changes**

- Removed `data_checks` parameter, `data_check_results` and data checks logic from `AutoMLSearch` #1935
- Deleted `random_state` argument #1985
- Updated `Woodwork` version requirement to `v0.0.11` #1996

- Documentation Changes

- **Testing Changes**

- Removed `build_docs` CI job in favor of RTD GH builder [#1974](#)
- Added tests to confirm support for Python 3.9 [#1724](#)
- Added tests to support Dask AutoML/Engine [#1990](#)
- Changed `build_conda_pkg` job to use `latest_release_changes` branch in the feed-stock. [#1979](#)

Warning:**Breaking Changes**

- Changed `AutoMLSearch` to default `optimize_thresholds` to `True` [#1943](#)
- Removed `data_checks` parameter, `data_check_results` and data checks logic from `AutoMLSearch`. To run the data checks which were previously run by default in `AutoMLSearch`, please call `DefaultDataChecks().validate(X_train, y_train)` or take a look at our documentation for more examples. [#1935](#)
- Deleted `random_state` argument [#1985](#)

v0.20.0 Mar. 10, 2021

- **Enhancements**

- Added a GitHub Action for Detecting dependency changes [#1933](#)
- Create a separate CV split to train stacked ensembler on for `AutoMLSearch` [#1814](#)
- Added a GitHub Action for Linux unit tests [#1846](#)
- Added `ARIMAREgressor` estimator [#1894](#)
- Added `DataCheckAction` class and `DataCheckActionCode` enum [#1896](#)
- Updated Woodwork requirement to `v0.0.10` [#1900](#)
- Added `BalancedClassificationDataCVSplit` and `BalancedClassificationDataTVSplit` to `AutoMLSearch` [#1875](#)
- Update default classification data splitter to use downsampling for highly imbalanced data [#1875](#)
- Updated `describe_pipeline` to return more information, including `id` of pipelines used for ensemble models [#1909](#)
- Added utility method to create list of components from a list of `DataCheckAction` [#1907](#)
- Updated `validate` method to include a `action` key in returned dictionary for all `DataCheck` and `DataChecks` [#1916](#)
- Aggregating the shap values for predictions that we know the provenance of, e.g. OHE, text, and date-time. [#1901](#)
- Improved error message when custom objective is passed as a string in `pipeline.score` [#1941](#)
- Added `score_pipelines` and `train_pipelines` methods to `AutoMLSearch` [#1913](#)
- Added support for pandas version 1.2.0 [#1708](#)
- Added `score_batch` and `train_batch` abstract methods to `EngineBase` and implementations in `SequentialEngine` [#1913](#)

- Added ability to handle index columns in `AutoMLSearch` and `DataChecks` #2138
- **Fixes**
 - Removed CI check for `check_dependencies_updated_linux` #1950
 - Added metaclass for time series pipelines and fix binary classification pipeline `predict` not using objective if it is passed as a named argument #1874
 - Fixed stack trace in prediction explanation functions caused by mixed string/numeric pandas column names #1871
 - Fixed stack trace caused by passing pipelines with duplicate names to `AutoMLSearch` #1932
 - Fixed `AutoMLSearch.get_pipelines` returning pipelines with the same attributes #1958
- **Changes**
 - Reversed GitHub Action for Linux unit tests until a fix for report generation is found #1920
 - Updated `add_results` in `AutoMLAlgorithm` to take in entire pipeline results dictionary from `AutoMLSearch` #1891
 - Updated `ClassImbalanceDataCheck` to look for severe class imbalance scenarios #1905
 - Deleted the `explain_prediction` function #1915
 - Removed `HighVarianceCVDataCheck` and converted it to an `AutoMLSearch` method instead #1928
 - Removed warning in `InvalidTargetDataCheck` returned when numeric binary classification targets are not (0, 1) #1959
- **Documentation Changes**
 - Updated `model_understanding.ipynb` to demo the two-way partial dependence capability #1919
- **Testing Changes**

Warning:**Breaking Changes**

- Deleted the `explain_prediction` function #1915
- Removed `HighVarianceCVDataCheck` and converted it to an `AutoMLSearch` method instead #1928
- Added `score_batch` and `train_batch` abstract methods to `EngineBase`. These need to be implemented in `Engine` subclasses #1913

v0.19.0 Feb. 23, 2021

- **Enhancements**
 - Added a GitHub Action for Python windows unit tests #1844
 - Added a GitHub Action for checking updated release notes #1849
 - Added a GitHub Action for Python lint checks #1837
 - Adjusted `explain_prediction`, `explain_predictions` and `explain_predictions_best_worst` to handle timeseries problems. #1818

- Updated `InvalidTargetDataCheck` to check for mismatched indices in target and features [#1816](#)
- Updated `Woodwork` structures returned from components to support `Woodwork` logical type overrides set by the user [#1784](#)
- Updated estimators to keep track of input feature names during `fit()` [#1794](#)
- Updated `visualize_decision_tree` to include feature names in output [#1813](#)
- Added `is_bounded_like_percentage` property for objectives. If true, the `calculate_percent_difference` method will return the absolute difference rather than relative difference [#1809](#)
- Added full error traceback to `AutoMLSearch` logger file [#1840](#)
- Changed `TargetEncoder` to preserve custom indices in the data [#1836](#)
- Refactored `explain_predictions` and `explain_predictions_best_worst` to only compute features once for all rows that need to be explained [#1843](#)
- Added custom random undersampler data splitter for classification [#1857](#)
- Updated `OutliersDataCheck` implementation to calculate the probability of having no outliers [#1855](#)
- Added `Engines` pipeline processing API [#1838](#)
- **Fixes**
 - Changed `EngineBase` `random_state` arg to `random_seed` and same for user guide docs [#1889](#)
- **Changes**
 - Modified `calculate_percent_difference` so that division by 0 is now `inf` rather than `nan` [#1809](#)
 - Removed `text_columns` parameter from `LSA` and `TextFeaturizer` components [#1652](#)
 - Added `random_seed` as an argument to our `automl/pipeline/component` API. Using `random_state` will raise a warning [#1798](#)
 - Added `DataCheckError` message in `InvalidTargetDataCheck` if input target is `None` and removed exception raised [#1866](#)
- **Documentation Changes**
- **Testing Changes**
 - Added back coverage for `_get_feature_provenance` in `TextFeaturizer` after `text_columns` was removed [#1842](#)
 - Pin `graphviz` version for windows builds [#1847](#)
 - Unpin `graphviz` version for windows builds [#1851](#)

Warning:**Breaking Changes**

- Added a deprecation warning to `explain_prediction`. It will be deleted in the next release. [#1860](#)

v0.18.2 Feb. 10, 2021

- **Enhancements**

- Added uniqueness score data check #1785
- Added “dataframe” output format for prediction explanations #1781
- Updated LightGBM estimators to handle `pandas.MultiIndex` #1770
- Sped up permutation importance for some pipelines #1762
- Added sparsity data check #1797
- Confirmed support for threshold tuning for binary time series classification problems #1803
- Fixes
- Changes
- **Documentation Changes**
 - Added section on conda to the contributing guide #1771
 - Updated release process to reflect freezing *main* before perf tests #1787
 - Moving some prs to the right section of the release notes #1789
 - Tweak README.md. #1800
 - Fixed back arrow on install page docs #1795
 - Fixed docstring for *ClassImbalanceDataCheck.validate()* #1817
- Testing Changes

v0.18.1 Feb. 1, 2021

- **Enhancements**
 - Added `graph_t_sne` as a visualization tool for high dimensional data #1731
 - Added the ability to see the linear coefficients of features in linear models terms #1738
 - Added support for `scikit-learn v0.24.0` #1733
 - Added support for `scipy v1.6.0` #1752
 - Added SVM Classifier and Regressor to estimators #1714 #1761
- **Fixes**
 - Addressed bug with `partial_dependence` and categorical data with more categories than grid resolution #1748
 - Removed `random_state` arg from `get_pipelines` in `AutoMLSearch` #1719
 - Pinned `pyzmq` at less than 22.0.0 till we add support #1756
- **Changes**
 - Updated components and pipelines to return `Woodwork` data structures #1668
 - Updated `clone()` for pipelines and components to copy over random state automatically #1753
 - Dropped support for Python version 3.6 #1751
 - Removed deprecated `verbose` flag from `AutoMLSearch` parameters #1772
- **Documentation Changes**
 - Add Twitter and Github link to documentation toolbar #1754
 - Added Open Graph info to documentation #1758

- Testing Changes

Warning:

Breaking Changes

- Components and pipelines return `Woodwork` data structures instead of `pandas` data structures [#1668](#)
- Python 3.6 will not be actively supported due to discontinued support from EvalML dependencies.
- Deprecated `verbose` flag is removed for `AutoMLSearch` [#1772](#)

v0.18.0 Jan. 26, 2021

- **Enhancements**

- Added RMSLE, MSLE, and MAPE to core objectives while checking for negative target values in `invalid_targets_data_check` [#1574](#)
- Added validation checks for binary problems with regression-like datasets and multiclass problems without true multiclass targets in `invalid_targets_data_check` [#1665](#)
- Added time series support for `make_pipeline` [#1566](#)
- Added target name for output of pipeline `predict` method [#1578](#)
- Added multiclass check to `InvalidTargetDataCheck` for two examples per class [#1596](#)
- Added support for `graphviz` v0.16 [#1657](#)
- Enhanced time series pipelines to accept empty features [#1651](#)
- Added KNN Classifier to estimators. [#1650](#)
- Added support for list inputs for objectives [#1663](#)
- Added support for `AutoMLSearch` to handle time series classification pipelines [#1666](#)
- Enhanced `DelayedFeaturesTransformer` to encode categorical features and targets before delaying them [#1691](#)
- Added 2-way dependence plots. [#1690](#)
- Added ability to directly iterate through components within Pipelines [#1583](#)

- **Fixes**

- Fixed inconsistent attributes and added Exceptions to docs [#1673](#)
- Fixed `TargetLeakageDataCheck` to use `Woodwork` `mutual_information` rather than using `Pandas`’ `Pearson Correlation` [#1616](#)
- Fixed thresholding for pipelines in `AutoMLSearch` to only threshold binary classification pipelines [#1622](#) [#1626](#)
- Updated `load_data` to return `Woodwork` structures and update default parameter value for `index` to `None` [#1610](#)
- Pinned `scipy` at < 1.6.0 while we work on adding support [#1629](#)
- Fixed data check message formatting in `AutoMLSearch` [#1633](#)
- Addressed stacked ensemble component for `scikit-learn` v0.24 support by setting `shuffle=True` for default CV [#1613](#)
- Fixed bug where `Imputer` reset the index on `X` [#1590](#)

- Fixed `AutoMLSearch` `stacktrace` when a custom objective was passed in as a primary objective or additional objective [#1575](#)
- Fixed custom index bug for MAPE objective [#1641](#)
- Fixed index bug for `TextFeaturizer` and LSA components [#1644](#)
- Limited `load_fraud` dataset loaded into `automl.ipynb` [#1646](#)
- `add_to_rankings` updates `AutoMLSearch.best_pipeline` when necessary [#1647](#)
- Fixed bug where time series baseline estimators were not receiving `gap` and `max_delay` in `AutoMLSearch` [#1645](#)
- Fixed jupyter notebooks to help the RTD buildtime [#1654](#)
- Added `positive_only` objectives to `non_core_objectives` [#1661](#)
- Fixed stacking argument `n_jobs` for `IterativeAlgorithm` [#1706](#)
- Updated CatBoost estimators to return self in `.fit()` rather than the underlying model for consistency [#1701](#)
- Added ability to initialize pipeline parameters in `AutoMLSearch` constructor [#1676](#)

- **Changes**

- Added labeling to `graph_confusion_matrix` [#1632](#)
- Rerunning search for `AutoMLSearch` results in a message thrown rather than failing the search, and removed `has_searched` property [#1647](#)
- Changed tuner class to allow and ignore single parameter values as input [#1686](#)
- Capped LightGBM version limit to remove bug in docs [#1711](#)
- Removed support for `np.random.RandomState` in EvalML [#1727](#)

- **Documentation Changes**

- Update Model Understanding in the user guide to include `visualize_decision_tree` [#1678](#)
- Updated docs to include information about `AutoMLSearch` callback parameters and methods [#1577](#)
- Updated docs to prompt users to install `graphviz` on Mac [#1656](#)
- Added `infer_feature_types` to the `start.ipynb` guide [#1700](#)
- Added multicollinearity data check to API reference and docs [#1707](#)

- **Testing Changes**

Warning:
Breaking Changes

- Removed `has_searched` property from `AutoMLSearch` [#1647](#)
- Components and pipelines return `Woodwork` data structures instead of `pandas` data structures [#1668](#)
- Removed support for `np.random.RandomState` in EvalML. Rather than passing `np.random.RandomState` as component and pipeline `random_state` values, we use `int random_seed` [#1727](#)

- **Enhancements**

- Added `save_plot` that allows for saving figures from different backends #1588
- Added `LightGBM Regressor` to regression components #1459
- Added `visualize_decision_tree` for tree visualization with `decision_tree_data_from_estimator` and `decision_tree_data_from_pipeline` to reformat tree structure output #1511
- Added *DFS Transformer* component into transformer components #1454
- Added MAPE to the standard metrics for time series problems and update objectives #1510
- Added `graph_prediction_vs_actual_over_time` and `get_prediction_vs_actual_over_time_data` to the model understanding module for time series problems #1483
- Added a `ComponentGraph` class that will support future pipelines as directed acyclic graphs #1415
- Updated data checks to accept Woodwork data structures #1481
- Added parameter to `InvalidTargetDataCheck` to show only top unique values rather than all unique values #1485
- Added multicollinearity data check #1515
- Added baseline pipeline and components for time series regression problems #1496
- Added more information to users about ensembling behavior in `AutoMLSearch` #1527
- Add woodwork support for more utility and graph methods #1544
- Changed `DateTimeFeaturizer` to encode features as int #1479
- Return trained pipelines from `AutoMLSearch.best_pipeline` #1547
- Added utility method so that users can set feature types without having to learn about Woodwork directly #1555
- Added Linear Discriminant Analysis transformer for dimensionality reduction #1331
- Added multiclass support for `partial_dependence` and `graph_partial_dependence` #1554
- Added `TimeSeriesBinaryClassificationPipeline` and `TimeSeriesMulticlassClassificationPipeline` classes #1528
- Added `make_data_splitter` method for easier automl data split customization #1568
- Integrated `ComponentGraph` class into Pipelines for full non-linear pipeline support #1543
- Update `AutoMLSearch` constructor to take training data instead of search and `add_to_leaderboard` #1597
- Update `split_data` helper args #1597
- Add problem type utils `is_regression`, `is_classification`, `is_timeseries` #1597
- Rename `AutoMLSearch` `data_split` arg to `data_splitter` #1569

- **Fixes**

- Fix `AutoML` not passing CV folds to `DefaultDataChecks` for usage by `ClassImbalanceDataCheck` #1619
- Fix Windows CI jobs: install numba via conda, required for shap #1490

- Added custom-index support for `reset-index-get_prediction_vs_actual_over_time_data` #1494
- Fix `generate_pipeline_code` to account for boolean and None differences between Python and JSON #1524 #1531
- Set max value for plotly and xgboost versions while we debug CI failures with newer versions #1532
- Undo version pinning for plotly #1533
- Fix ReadTheDocs build by updating the version of `setuptools` #1561
- Set `random_state` of data splitter in `AutoMLSearch` to take int to keep consistency in the resulting splits #1579
- Pin `sklearn` version while we work on adding support #1594
- Pin `pandas` at <1.2.0 while we work on adding support #1609
- Pin `graphviz` at <0.16 while we work on adding support #1609

- **Changes**

- Reverting `save_graph` #1550 to resolve kaleido build issues #1585
- Update circleci badge to apply to main #1489
- Added script to generate github markdown for releases #1487
- Updated selection using `pandas` dtypes to selecting using Woodwork logical types #1551
- Updated dependencies to fix `ImportError: cannot import name 'MaskedArray' from 'sklearn.utils.fixes'` error and to address Woodwork and Featuretool dependencies #1540
- Made `get_prediction_vs_actual_data()` a public method #1553
- Updated Woodwork version requirement to v0.0.7 #1560
- Move data splitters from `evalml.automl.data_splitters` to `evalml.preprocessing.data_splitters` #1597
- Rename “# Testing” in automl log output to “# Validation” #1597

- **Documentation Changes**

- Added partial dependence methods to API reference #1537
- Updated documentation for confusion matrix methods #1611

- **Testing Changes**

- Set `n_jobs=1` in most unit tests to reduce memory #1505

Warning:

Breaking Changes

- Updated minimal dependencies: `numpy>=1.19.1`, `pandas>=1.1.0`, `scikit-learn>=0.23.1`, `scikit-optimize>=0.8.1`
- Updated `AutoMLSearch.best_pipeline` to return a trained pipeline. Pass in `train_best_pipeline=False` to `AutoMLSearch` in order to return an untrained pipeline.
- Pipeline component instances can no longer be iterated through using `Pipeline.component_graph` #1543

- Update `AutoMLSearch` constructor to take training data instead of search and `add_to_leaderboard` #1597
- Update `split_data` helper args #1597
- Move data splitters from `evalml.automl.data_splitters` to `evalml.preprocessing.data_splitters` #1597
- Rename `AutoMLSearch` `data_split` arg to `data_splitter` #1569

v0.16.1 Dec. 1, 2020

- **Enhancements**
 - Pin `woodwork` version to `v0.0.6` to avoid breaking changes #1484
 - Updated `Woodwork` to `>=0.0.5` in `core-requirements.txt` #1473
 - Removed `copy_dataframe` parameter for `Woodwork`, updated `Woodwork` to `>=0.0.6` in `core-requirements.txt` #1478
 - Updated `detect_problem_type` to use `pandas.api.is_numeric_dtype` #1476
- **Changes**
 - Changed `make_clean` to delete coverage reports as a convenience for developers #1464
 - Set `n_jobs=-1` by default for stacked ensemble components #1472
- **Documentation Changes**
 - Updated pipeline and component documentation and demos to use `Woodwork` #1466
- **Testing Changes**
 - Update dependency update checker to use everything from core and optional dependencies #1480

v0.16.0 Nov. 24, 2020

- **Enhancements**
 - Updated pipelines and `make_pipeline` to accept `Woodwork` inputs #1393
 - Updated components to accept `Woodwork` inputs #1423
 - Added ability to freeze hyperparameters for `AutoMLSearch` #1284
 - Added `Target Encoder` into transformer components #1401
 - Added callback for error handling in `AutoMLSearch` #1403
 - Added the index id to the `explain_predictions_best_worst` output to help users identify which rows in their data are included #1365
 - The `top_k` features displayed in `explain_predictions_*` functions are now determined by the magnitude of shap values as opposed to the `top_k` largest and smallest shap values. #1374
 - Added a problem type for time series regression #1386
 - Added a `is_defined_for_problem_type` method to `ObjectiveBase` #1386
 - Added a `random_state` parameter to `make_pipeline_from_components` function #1411
 - Added `DelayedFeaturesTransformer` #1396
 - Added a `TimeSeriesRegressionPipeline` class #1418

- Removed `core-requirements.txt` from the package distribution #1429
- Updated data check messages to include a “*code*” and “*details*” fields #1451, #1462
- Added a `TimeSeriesSplit` data splitter for time series problems #1441
- Added a `problem_configuration` parameter to `AutoMLSearch` #1457
- **Fixes**
 - Fixed `IndexError` raised in `AutoMLSearch` when `ensembling = True` but only one pipeline to iterate over #1397
 - Fixed stacked ensemble input bug and `LightGBM` warning and bug in `AutoMLSearch` #1388
 - Updated enum classes to show possible enum values as attributes #1391
 - Updated calls to `Woodwork`’s `to_pandas()` to `to_series()` and `to_dataframe()` #1428
 - Fixed bug in OHE where column names were not guaranteed to be unique #1349
 - Fixed bug with percent improvement of `ExpVariance` objective on data with highly skewed target #1467
 - Fix `SimpleImputer` error which occurs when all features are bool type #1215
- **Changes**
 - Changed `OutliersDataCheck` to return the list of columns, rather than rows, that contain outliers #1377
 - Simplified and cleaned output for Code Generation #1371
 - Reverted changes from #1337 #1409
 - Updated data checks to return dictionary of warnings and errors instead of a list #1448
 - Updated `AutoMLSearch` to pass `Woodwork` data structures to every pipeline (instead of pandas `DataFrames`) #1450
 - Update `AutoMLSearch` to default to `max_batches=1` instead of `max_iterations=5` #1452
 - Updated `_evaluate_pipelines` to consolidate side effects #1410
- **Documentation Changes**
 - Added description of CLA to contributing guide, updated description of draft PRs #1402
 - Updated documentation to include all data checks, `DataChecks`, and usage of data checks in `AutoML` #1412
 - Updated docstrings from `np.array` to `np.ndarray` #1417
 - Added section on stacking ensembles in `AutoMLSearch` documentation #1425
- **Testing Changes**
 - Removed `category_encoders` from `test-requirements.txt` #1373
 - Tweak `codecov.io` settings again to avoid flakes #1413
 - Modified `make lint` to check notebook versions in the docs #1431
 - Modified `make lint-fix` to standardize notebook versions in the docs #1431
 - Use new version of pull request Github Action for dependency check (#1443)

- Reduced number of workers for tests to 4 [#1447](#)

Warning:

Breaking Changes

- The `top_k` and `top_k_features` parameters in `explain_predictions_*` functions now return `k` features as opposed to `2 * k` features [#1374](#)
- Renamed `problem_type` to `problem_types` in `RegressionObjective`, `BinaryClassificationObjective`, and `MulticlassClassificationObjective` [#1319](#)
- Data checks now return a dictionary of warnings and errors instead of a list [#1448](#)

v0.15.0 Oct. 29, 2020

• **Enhancements**

- Added stacked ensemble component classes (`StackedEnsembleClassifier`, `StackedEnsembleRegressor`) [#1134](#)
- Added stacked ensemble components to `AutoMLSearch` [#1253](#)
- Added `DecisionTreeClassifier` and `DecisionTreeRegressor` to `AutoML` [#1255](#)
- Added `graph_prediction_vs_actual` in `model_understanding` for regression problems [#1252](#)
- Added parameter to `OneHotEncoder` to enable filtering for features to encode for [#1249](#)
- Added `percent-better-than-baseline` for all objectives to `automl.results` [#1244](#)
- Added `HighVarianceCVDataCheck` and replaced synonymous warning in `AutoMLSearch` [#1254](#)
- Added *PCA Transformer* component for dimensionality reduction [#1270](#)
- Added `generate_pipeline_code` and `generate_component_code` to allow for code generation given a pipeline or component instance [#1306](#)
- Added *PCA Transformer* component for dimensionality reduction [#1270](#)
- Updated `AutoMLSearch` to support Woodwork data structures [#1299](#)
- Added `cv_folds` to `ClassImbalanceDataCheck` and added this check to `DefaultDataChecks` [#1333](#)
- Make `max_batches` argument to `AutoMLSearch.search` public [#1320](#)
- Added text support to `automl search` [#1062](#)
- Added `_pipelines_per_batch` as a private argument to `AutoMLSearch` [#1355](#)

• **Fixes**

- Fixed ML performance issue with ordered datasets: always shuffle data in `automl`'s default CV splits [#1265](#)
- Fixed broken `evalml info` CLI command [#1293](#)
- Fixed `boosting type='rf'` for `LightGBM Classifier`, as well as `num_leaves` error [#1302](#)
- Fixed bug in `explain_predictions_best_worst` where a custom index in the target variable would cause a `ValueError` [#1318](#)

- Added stacked ensemble estimators to `evalml.pipelines.__init__` file #1326
- Fixed bug in OHE where calls to transform were not deterministic if `top_n` was less than the number of categories in a column #1324
- Fixed LightGBM warning messages during AutoMLSearch #1342
- Fix warnings thrown during AutoMLSearch in `HighVarianceCVDDataCheck` #1346
- Fixed bug where `TrainingValidationSplit` would return invalid location indices for dataframes with a custom index #1348
- Fixed bug where the AutoMLSearch `random_state` was not being passed to the created pipelines #1321

- **Changes**

- Allow `add_to_rankings` to be called before AutoMLSearch is called #1250
- Removed Graphviz from test-requirements to add to requirements.txt #1327
- Removed `max_pipelines` parameter from AutoMLSearch #1264
- Include editable installs in all install make targets #1335
- Made pip dependencies *featuretools* and *nlp_primitives* core dependencies #1062
- Removed *PartOfSpeechCount* from *TextFeaturizer* transform primitives #1062
- Added warning for `partial_dependency` when the feature includes null values #1352

- **Documentation Changes**

- Fixed and updated code blocks in Release Notes #1243
- Added DecisionTree estimators to API Reference #1246
- Changed class inheritance display to flow vertically #1248
- Updated cost-benefit tutorial to use a holdout/test set #1159
- Added `evalml info` command to documentation #1293
- Miscellaneous doc updates #1269
- Removed conda pre-release testing from the release process document #1282
- Updates to contributing guide #1310
- Added Alteryx footer to docs with Twitter and Github link #1312
- Added documentation for evalml installation for Python 3.6 #1322
- Added documentation changes to make the API Docs easier to understand #1323
- Fixed documentation for `feature_importance` #1353
- Added tutorial for running *AutoML* with text data #1357
- Added documentation for woodwork integration with automl search #1361

- **Testing Changes**

- Added tests for `jupyter_check` to handle IPython #1256
- Cleaned up `make_pipeline` tests to test for all estimators #1257
- Added a test to check conda build after merge to main #1247
- Removed code that was lacking codecov for `__main__.py` and unnecessary #1293

- Codecov: round coverage up instead of down [#1334](#)
- Add DockerHub credentials to CI testing environment [#1356](#)
- Add DockerHub credentials to conda testing environment [#1363](#)

Warning:**Breaking Changes**

- Renamed `LabelLeakageDataCheck` to `TargetLeakageDataCheck` [#1319](#)
- `max_pipelines` parameter has been removed from `AutoMLSearch`. Please use `max_iterations` instead. [#1264](#)
- `AutoMLSearch.search()` will now log a warning if the input is not a Woodwork data structure (pandas, numpy) [#1299](#)
- Make `max_batches` argument to `AutoMLSearch.search` public [#1320](#)
- Removed unused argument `feature_types` from `AutoMLSearch.search` [#1062](#)

v0.14.1 Sep. 29, 2020**• Enhancements**

- Updated partial dependence methods to support calculating numeric columns in a dataset with non-numeric columns [#1150](#)
- Added `get_feature_names` on `OneHotEncoder` [#1193](#)
- Added `detect_problem_type` to `problem_type/utils.py` to automatically detect the problem type given targets [#1194](#)
- Added `LightGBM` to `AutoMLSearch` [#1199](#)
- Updated `scikit-learn` and `scikit-optimize` to use latest versions - 0.23.2 and 0.8.1 respectively [#1141](#)
- Added `__str__` and `__repr__` for pipelines and components [#1218](#)
- Included internal target check for both training and validation data in `AutoMLSearch` [#1226](#)
- Added `ProblemTypes.all_problem_types` helper to get list of supported problem types [#1219](#)
- Added `DecisionTreeClassifier` and `DecisionTreeRegressor` classes [#1223](#)
- Added `ProblemTypes.all_problem_types` helper to get list of supported problem types [#1219](#)
- `DataChecks` can now be parametrized by passing a list of `DataCheck` classes and a parameter dictionary [#1167](#)
- Added first CV fold score as validation score in `AutoMLSearch.rankings` [#1221](#)
- Updated `flake8` configuration to enable linting on `__init__.py` files [#1234](#)
- Refined `make_pipeline_from_components` implementation [#1204](#)

• Fixes

- Updated GitHub URL after migration to Alteryx GitHub org [#1207](#)
- Changed Problem Type enum to be more similar to the string name [#1208](#)

- Wrapped call to scikit-learn’s partial dependence method in a `try/finally` block [#1232](#)

- **Changes**

- Added `allow_writing_files` as a named argument to CatBoost estimators. [#1202](#)
- Added `solver` and `multi_class` as named arguments to `LogisticRegressionClassifier` [#1202](#)
- Replaced pipeline’s `._transform` method to evaluate all the preprocessing steps of a pipeline with `.compute_estimator_features` [#1231](#)
- Changed default large dataset train/test splitting behavior [#1205](#)

- **Documentation Changes**

- Included description of how to access the component instances and features for pipeline user guide [#1163](#)
- Updated API docs to refer to target as “target” instead of “labels” for non-classification tasks and minor docs cleanup [#1160](#)
- Added Class Imbalance Data Check to `api_reference.rst` [#1190](#) [#1200](#)
- Added pipeline properties to API reference [#1209](#)
- Clarified what the objective parameter in AutoML is used for in AutoML API reference and AutoML user guide [#1222](#)
- Updated API docs to include `skopt.space.Categorical` option for component hyperparameter range definition [#1228](#)
- Added install documentation for `libomp` in order to use LightGBM on Mac [#1233](#)
- Improved description of `max_iterations` in documentation [#1212](#)
- Removed unused code from sphinx conf [#1235](#)

- **Testing Changes**

Warning:

Breaking Changes

- `DefaultDataChecks` now accepts a `problem_type` parameter that must be specified [#1167](#)
- Pipeline’s `._transform` method to evaluate all the preprocessing steps of a pipeline has been replaced with `.compute_estimator_features` [#1231](#)
- `get_objectives` has been renamed to `get_core_objectives`. This function will now return a list of valid objective instances [#1230](#)

v0.13.2 Sep. 17, 2020

- **Enhancements**

- Added `output_format` field to explain predictions functions [#1107](#)
- Modified `get_objective` and `get_objectives` to be able to return any objective in `evalml.objectives` [#1132](#)
- Added a `return_instance` boolean parameter to `get_objective` [#1132](#)
- Added `ClassImbalanceDataCheck` to determine whether target imbalance falls below a given threshold [#1135](#)

- Added label encoder to LightGBM for binary classification #1152
- Added labels for the row index of confusion matrix #1154
- Added `AutoMLSearch` object as another parameter in search callbacks #1156
- Added the corresponding probability threshold for each point displayed in `graph_roc_curve` #1161
- Added `__eq__` for `ComponentBase` and `PipelineBase` #1178
- Added support for multiclass classification for `roc_curve` #1164
- Added `categories` accessor to `OneHotEncoder` for listing the categories associated with a feature #1182
- Added utility function to create pipeline instances from a list of component instances #1176
- **Fixes**
 - Fixed XGBoost column names for partial dependence methods #1104
 - Removed dead code validating column type from `TextFeaturizer` #1122
 - Fixed issue where `Imputer` cannot fit when there is `None` in a categorical or boolean column #1144
 - `OneHotEncoder` preserves the custom index in the input data #1146
 - Fixed representation for `ModelFamily` #1165
 - Removed duplicate `nbsphinx` dependency in `dev-requirements.txt` #1168
 - Users can now pass in any valid kwargs to all estimators #1157
 - Remove broken accessor `OneHotEncoder.get_feature_names` and unneeded base class #1179
 - Removed LightGBM Estimator from AutoML models #1186
- **Changes**
 - Pinned `scikit-optimize` version to 0.7.4 #1136
 - Removed `tqdm` as a dependency #1177
 - Added `lightgbm` version 3.0.0 to `latest_dependency_versions.txt` #1185
 - Rename `max_pipelines` to `max_iterations` #1169
- **Documentation Changes**
 - Fixed API docs for `AutoMLSearch.add_result_callback` #1113
 - Added a step to our release process for pushing our latest version to conda-forge #1118
 - Added warning for missing `ipywidgets` dependency for using `PipelineSearchPlots` on Jupyterlab #1145
 - Updated `README.md` example to load demo dataset #1151
 - Swapped mapping of breast cancer targets in `model_understanding.ipynb` #1170
- **Testing Changes**
 - Added test confirming `TextFeaturizer` never outputs null values #1122
 - Changed Python version of `Update Dependencies` action to 3.8.x #1137
 - Fixed release notes check-in test for `Update Dependencies` actions #1172

Warning:**Breaking Changes**

- `get_objective` will now return a class definition rather than an instance by default [#1132](#)
- Deleted `OPTIONS` dictionary in `evalml.objectives.utils.py` [#1132](#)
- If specifying an objective by string, the string must now match the objective's name field, case-insensitive [#1132](#)
- Passing “Cost Benefit Matrix”, “Fraud Cost”, “Lead Scoring”, “Mean Squared Log Error”, “Recall”, “Recall Macro”, “Recall Micro”, “Recall Weighted”, or “Root Mean Squared Log Error” to `AutoMLSearch` will now result in a `ValueError` rather than an `ObjectiveNotFoundError` [#1132](#)
- Search callbacks `start_iteration_callback` and `add_results_callback` have changed to include a copy of the `AutoMLSearch` object as a third parameter [#1156](#)
- Deleted `OneHotEncoder.get_feature_names` method which had been broken for a while, in favor of pipelines' `input_feature_names` [#1179](#)
- Deleted empty base class `CategoricalEncoder` which `OneHotEncoder` component was inheriting from [#1176](#)
- Results from `roc_curve` will now return as a list of dictionaries with each dictionary representing a class [#1164](#)
- `max_pipelines` now raises a `DeprecationWarning` and will be removed in the next release. `max_iterations` should be used instead. [#1169](#)

v0.13.1 Aug. 25, 2020• **Enhancements**

- Added Cost-Benefit Matrix objective for binary classification [#1038](#)
- Split `fill_value` into `categorical_fill_value` and `numeric_fill_value` for `Imputer` [#1019](#)
- Added `explain_predictions` and `explain_predictions_best_worst` for explaining multiple predictions with SHAP [#1016](#)
- Added new LSA component for text featurization [#1022](#)
- Added guide on installing with conda [#1041](#)
- Added a “cost-benefit curve” util method to graph cost-benefit matrix scores vs. binary classification thresholds [#1081](#)
- Standardized error when calling `transform/predict` before `fit` for pipelines [#1048](#)
- Added `percent_better_than_baseline` to `AutoML` search rankings and full rankings table [#1050](#)
- Added one-way partial dependence and partial dependence plots [#1079](#)
- Added “Feature Value” column to prediction explanation reports. [#1064](#)
- Added LightGBM classification estimator [#1082](#), [#1114](#)
- Added `max_batches` parameter to `AutoMLSearch` [#1087](#)

• **Fixes**

- Updated `TextFeaturizer` component to no longer require an internet connection to run #1022
- Fixed non-deterministic element of `TextFeaturizer` transformations #1022
- Added a `StandardScaler` to all `ElasticNet` pipelines #1065
- Updated cost-benefit matrix to normalize score #1099
- Fixed logic in `calculate_percent_difference` so that it can handle negative values #1100

- **Changes**

- Added `needs_fitting` property to `ComponentBase` #1044
- Updated references to data types to use datatype lists defined in `evalml.utils.gen_utils` #1039
- Remove maximum version limit for SciPy dependency #1051
- Moved `all_components` and other component importers into runtime methods #1045
- Consolidated graphing utility methods under `evalml.utils.graph_utils` #1060
- Made slight tweaks to how `TextFeaturizer` uses `featuretools`, and did some refactoring of that and of LSA #1090
- Changed `show_all_features` parameter into `importance_threshold`, which allows for thresholding feature importance #1097, #1103

- **Documentation Changes**

- Update `setup.py` URL to point to the github repo #1037
- Added tutorial for using the cost-benefit matrix objective #1088
- Updated `model_understanding.ipynb` to include documentation for using plotly on Jupyter Lab #1108

- **Testing Changes**

- Refactor CircleCI tests to use matrix jobs (#1043)
- Added a test to check that all test directories are included in evalml package #1054

Warning:**Breaking Changes**

- `confusion_matrix` and `normalize_confusion_matrix` have been moved to `evalml.utils` #1038
- All graph utility methods previously under `evalml.pipelines.graph_utils` have been moved to `evalml.utils.graph_utils` #1060

v0.12.2 Aug. 6, 2020

- **Enhancements**

- Add save/load method to components #1023
- Expose `pickle_protocol` as optional arg to save/load #1023
- Updated estimators used in AutoML to include `ExtraTrees` and `ElasticNet` estimators #1030

- **Fixes**

- **Changes**
 - Removed `DeprecationWarning` for `SimpleImputer` [#1018](#)
- **Documentation Changes**
 - Add note about version numbers to release process docs [#1034](#)
- **Testing Changes**
 - Test files are now included in the evalml package [#1029](#)

v0.12.0 Aug. 3, 2020

- **Enhancements**
 - Added string and categorical targets support for binary and multiclass pipelines and check for numeric targets for `DetectLabelLeakage` data check [#932](#)
 - Added clear exception for regression pipelines if target datatype is string or categorical [#960](#)
 - Added target column names and class labels in `predict` and `predict_proba` output for pipelines [#951](#)
 - Added `_compute_shap_values` and `normalize_values` to `pipelines/explanations` module [#958](#)
 - Added `explain_prediction` feature which explains single predictions with SHAP [#974](#)
 - Added `Imputer` to allow different imputation strategies for numerical and categorical dtypes [#991](#)
 - Added support for configuring logfile path using env var, and don't create logger if there are filesystem errors [#975](#)
 - Updated catboost estimators' default parameters and automl hyperparameter ranges to speed up fit time [#998](#)
- **Fixes**
 - Fixed `ReadtheDocs` warning failure regarding embedded gif [#943](#)
 - Removed incorrect parameter passed to pipeline classes in `_add_baseline_pipelines` [#941](#)
 - Added universal error for calling `predict`, `predict_proba`, `transform`, and `feature_importances` before fitting [#969](#), [#994](#)
 - Made `TextFeaturizer` component and pip dependencies `featuretools` and `nlp_primitives` optional [#976](#)
 - Updated imputation strategy in automl to no longer limit impute strategy to `most_frequent` for all features if there are any categorical columns [#991](#)
 - Fixed `UnboundLocalError` for `cv_pipeline` when automl search errors [#996](#)
 - Fixed `Imputer` to reset dataframe index to preserve behavior expected from `SimpleImputer` [#1009](#)
- **Changes**
 - Moved `get_estimators` to `evalml.pipelines.components.utils` [#934](#)
 - Modified Pipelines to raise `PipelineScoreError` when they encounter an error during scoring [#936](#)
 - Moved `evalml.model_families.list_model_families` to `evalml.pipelines.components.allowed_model_families` [#959](#)

- Renamed `DateTimeFeaturization` to `DateTimeFeaturizer` [#977](#)
- Added check to stop search and raise an error if all pipelines in a batch return NaN scores [#1015](#)
- **Documentation Changes**
 - Updated `README.md` [#963](#)
 - Reworded message when errors are returned from data checks in search [#982](#)
 - Added section on understanding model predictions with `explain_prediction` to User Guide [#981](#)
 - Added a section to the user guide and api reference about how XGBoost and CatBoost are not fully supported. [#992](#)
 - Added custom components section in user guide [#993](#)
 - Updated FAQ section formatting [#997](#)
 - Updated release process documentation [#1003](#)
- **Testing Changes**
 - Moved `predict_proba` and `predict` tests regarding string / categorical targets to `test_pipelines.py` [#972](#)
 - Fixed dependency update bot by updating python version to 3.7 to avoid frequent github version updates [#1002](#)

Warning:

Breaking Changes

- `get_estimators` has been moved to `evalml.pipelines.components.utils` (previously was under `evalml.pipelines.utils`) [#934](#)
- Removed the `raise_errors` flag in AutoML search. All errors during pipeline evaluation will be caught and logged. [#936](#)
- `evalml.model_families.list_model_families` has been moved to `evalml.pipelines.components.allowed_model_families` [#959](#)
- `TextFeaturizer`: the `featuretools` and `nlp_primitives` packages must be installed after installing evalml in order to use this component [#976](#)
- Renamed `DateTimeFeaturization` to `DateTimeFeaturizer` [#977](#)

v0.11.2 July 16, 2020

- **Enhancements**
 - Added `NoVarianceDataCheck` to `DefaultDataChecks` [#893](#)
 - Added text processing and featurization component `TextFeaturizer` [#913](#), [#924](#)
 - Added additional checks to `InvalidTargetDataCheck` to handle invalid target data types [#929](#)
 - `AutoMLSearch` will now handle `KeyboardInterrupt` and prompt user for confirmation [#915](#)
- **Fixes**
 - Makes `automl` results a read-only property [#919](#)

- **Changes**

- Deleted static pipelines and refactored tests involving static pipelines, removed `all_pipelines()` and `get_pipelines()` #904
- Moved `list_model_families` to `evalml.model_family.utils` #903
- Updated `all_pipelines`, `all_estimators`, `all_components` to use the same mechanism for dynamically generating their elements #898
- Rename master branch to main #918
- Add pypi release github action #923
- Updated `AutoMLSearch.search` stdout output and logging and removed tqdm progress bar #921
- Moved automl config checks previously in `search()` to `init` #933

- **Documentation Changes**

- Reorganized and rewrote documentation #937
- Updated to use pydata sphinx theme #937
- Updated docs to use `release_notes` instead of `changelog` #942

- **Testing Changes**

- Cleaned up fixture names and usages in tests #895

Warning:
Breaking Changes

- `list_model_families` has been moved to `evalml.model_family.utils` (previously was under `evalml.pipelines.utils`) #903
- `get_estimators` has been moved to `evalml.pipelines.components.utils` (previously was under `evalml.pipelines.utils`) #934
- Static pipeline definitions have been removed, but similar pipelines can still be constructed via creating an instance of `PipelineBase` #904
- `all_pipelines()` and `get_pipelines()` utility methods have been removed #904

v0.11.0 June 30, 2020

- **Enhancements**

- Added multiclass support for ROC curve graphing #832
- Added preprocessing component to drop features whose percentage of NaN values exceeds a specified threshold #834
- Added data check to check for problematic target labels #814
- Added `PerColumnImputer` that allows imputation strategies per column #824
- Added transformer to drop specific columns #827
- Added support for categories, `handle_error`, and `drop` parameters in `OneHotEncoder` #830 #897
- Added preprocessing component to handle `DateTime` columns featurization #838

- Added ability to clone pipelines and components #842
- Define getter method for component parameters #847
- Added utility methods to calculate and graph permutation importances #860, #880
- Added new utility functions necessary for generating dynamic preprocessing pipelines #852
- Added kwargs to all components #863
- Updated `AutoSearchBase` to use dynamically generated preprocessing pipelines #870
- Added `SelectColumns` transformer #873
- Added ability to evaluate additional pipelines for automl search #874
- Added `default_parameters` class property to components and pipelines #879
- Added better support for disabling data checks in automl search #892
- Added ability to save and load AutoML objects to file #888
- Updated `AutoSearchBase.get_pipelines` to return an untrained pipeline instance #876
- Saved learned binary classification thresholds in automl results cv data dict #876
- **Fixes**
 - Fixed bug where `SimpleImputer` cannot handle dropped columns #846
 - Fixed bug where `PerColumnImputer` cannot handle dropped columns #855
 - Enforce requirement that builtin components save all inputted values in their parameters dict #847
 - Don't list base classes in `all_components` output #847
 - Standardize all components to output pandas data structures, and accept either pandas or numpy #853
 - Fixed rankings and `full_rankings` error when search has not been run #894
- **Changes**
 - Update `all_pipelines` and `all_components` to try initializing pipelines/components, and on failure exclude them #849
 - Refactor `handle_components` to `handle_components_class`, standardize to `ComponentBase` subclass instead of instance #850
 - Refactor “blacklist”/“whitelist” to “allow”/“exclude” lists #854
 - Replaced `AutoClassificationSearch` and `AutoRegressionSearch` with `AutoMLSearch` #871
 - Renamed `feature_importances` and `permutation_importances` methods to use singular names (`feature_importance` and `permutation_importance`) #883
 - Updated automl default data splitter to train/validation split for large datasets #877
 - Added open source license, update some repo metadata #887
 - Removed dead code in `_get_preprocessing_components` #896
- **Documentation Changes**
 - Fix some typos and update the EvalML logo #872
- **Testing Changes**
 - Update the changelog check job to expect the new branching pattern for the deps update bot #836

- Check that all components output pandas datastructures, and can accept either pandas or numpy #853
- Replaced `AutoClassificationSearch` and `AutoRegressionSearch` with `AutoMLSearch` #871

Warning:**Breaking Changes**

- Pipelines' static `component_graph` field must contain either `ComponentBase` subclasses or `str`, instead of `ComponentBase` subclass instances #850
- Rename `handle_component` to `handle_component_class`. Now standardizes to `ComponentBase` subclasses instead of `ComponentBase` subclass instances #850
- Renamed `automl`'s `cv` argument to `data_split` #877
- Pipelines' and classifiers' `feature_importances` is renamed `feature_importance`, `graph_feature_importances` is renamed `graph_feature_importance` #883
- Passing `data_checks=None` to `automl` search will not perform any data checks as opposed to default checks. #892
- Pipelines to search for in AutoML are now determined automatically, rather than using the statically-defined pipeline classes. #870
- Updated `AutoSearchBase.get_pipelines` to return an untrained pipeline instance, instead of one which happened to be trained on the final cross-validation fold #876

v0.10.0 May 29, 2020• **Enhancements**

- Added baseline models for classification and regression, add functionality to calculate baseline models before searching in AutoML #746
- Port over highly-null guardrail as a data check and define `DefaultDataChecks` and `DisableDataChecks` classes #745
- Update `Tuner` classes to work directly with pipeline parameters dicts instead of flat parameter lists #779
- Add Elastic Net as a pipeline option #812
- Added new Pipeline option `ExtraTrees` #790
- Added precision-recall curve metrics and plot for binary classification problems in `evalml.pipeline.graph_utils` #794
- Update the default `automl` algorithm to search in batches, starting with default parameters for each pipeline and iterating from there #793
- Added `AutoMLAlgorithm` class and `IterativeAlgorithm` impl, separated from `AutoSearchBase` #793

• **Fixes**

- Update pipeline score to return nan score for any objective which throws an exception during scoring #787
- Fixed bug introduced in #787 where binary classification metrics requiring predicted probabilities error in scoring #798

- CatBoost and XGBoost classifiers and regressors can no longer have a learning rate of 0 [#795](#)

- **Changes**

- Cleanup pipeline `score` code, and cleanup codecov [#711](#)
- Remove `pass` for abstract methods for codecov [#730](#)
- Added `__str__` for AutoSearch object [#675](#)
- Add util methods to graph ROC and confusion matrix [#720](#)
- Refactor AutoBase to AutoSearchBase [#758](#)
- Updated AutoBase with `data_checks` parameter, removed previous `detect_label_leakage` parameter, and added functionality to run data checks before search in AutoML [#765](#)
- Updated our logger to use Python’s logging utils [#763](#)
- Refactor most of AutoSearchBase.`_do_iteration impl` into AutoSearchBase.`_evaluate` [#762](#)
- Port over all guardrails to use the new DataCheck API [#789](#)
- Expanded `import_or_raise` to catch all exceptions [#759](#)
- Adds RMSE, MSLE, RMSLE as standard metrics [#788](#)
- Don’t allow Recall to be used as an objective for AutoML [#784](#)
- Removed feature selection from pipelines [#819](#)
- Update default estimator parameters to make automl search faster and more accurate [#793](#)

- **Documentation Changes**

- Add instructions to freeze `master` on `release.md` [#726](#)
- Update release instructions with more details [#727](#) [#733](#)
- Add objective base classes to API reference [#736](#)
- Fix components API to match other modules [#747](#)

- **Testing Changes**

- Delete codecov `yml`, use codecov.io’s default [#732](#)
- Added unit tests for fraud cost, lead scoring, and standard metric objectives [#741](#)
- Update codecov client [#782](#)
- Updated AutoBase `__str__` test to include no parameters case [#783](#)
- Added unit tests for ExtraTrees pipeline [#790](#)
- If codecov fails to upload, fail build [#810](#)
- Updated Python version of dependency action [#816](#)
- Update the dependency update bot to use a suffix when creating branches [#817](#)

Warning:

Breaking Changes

- The `detect_label_leakage` parameter for AutoML classes has been removed and replaced by a `data_checks` parameter [#765](#)
- Moved ROC and confusion matrix methods from `evalml.pipeline.plot_utils` to `evalml.pipeline.graph_utils` [#720](#)
- Tuner classes require a pipeline hyperparameter range dict as an init arg instead of a space definition [#779](#)
- `Tuner.propose` and `Tuner.add` work directly with pipeline parameters dicts instead of flat parameter lists [#779](#)
- `PipelineBase.hyperparameters` and `custom_hyperparameters` use pipeline parameters dict format instead of being represented as a flat list [#779](#)
- All guardrail functions previously under `evalml.guardrails.utils` will be removed and replaced by data checks [#789](#)
- Recall disallowed as an objective for AutoML [#784](#)
- `AutoSearchBase` parameter `tuner` has been renamed to `tuner_class` [#793](#)
- `AutoSearchBase` parameter `possible_pipelines` and `possible_model_families` have been renamed to `allowed_pipelines` and `allowed_model_families` [#793](#)

v0.9.0 Apr. 27, 2020

• Enhancements

- Added `Accuracy` as a standard objective [#624](#)
- Added `verbose` parameter to `load_fraud` [#560](#)
- Added `Balanced Accuracy` metric for binary, multiclass [#612](#) [#661](#)
- Added `XGBoost` regressor and `XGBoost` regression pipeline [#666](#)
- Added `Accuracy` metric for multiclass [#672](#)
- Added objective name in `AutoBase.describe_pipeline` [#686](#)
- Added `DataCheck` and `DataChecks`, `Message` classes and relevant subclasses [#739](#)

• Fixes

- Removed direct access to `cls.component_graph` [#595](#)
- Add testing files to `.gitignore` [#625](#)
- Remove circular dependencies from `Makefile` [#637](#)
- Add error case for `normalize_confusion_matrix()` [#640](#)
- Fixed `XGBoostClassifier` and `XGBoostRegressor` bug with feature names that contain `[,]`, or `<` [#659](#)
- Update `make_pipeline_graph` to not accidentally create empty file when testing if path is valid [#649](#)
- Fix pip installation warning about `docsutils` version, from `boto` dependency [#664](#)
- Removed zero division warning for F1/precision/recall metrics [#671](#)
- Fixed `summary` for pipelines without estimators [#707](#)

• Changes

- Updated default objective for binary/multiclass classification to log loss #613
- Created classification and regression pipeline subclasses and removed objective as an attribute of pipeline classes #405
- Changed the output of `score` to return one dictionary #429
- Created binary and multiclass objective subclasses #504
- Updated objectives API #445
- Removed call to `get_plot_data` from AutoML #615
- Set `raise_error` to default to True for AutoML classes #638
- Remove unnecessary “u” prefixes on some unicode strings #641
- Changed one-hot encoder to return uint8 dtypes instead of ints #653
- Pipeline `_name` field changed to `custom_name` #650
- Removed `graphs.py` and moved methods into `PipelineBase` #657, #665
- Remove `s3fs` as a dev dependency #664
- Changed requirements-parser to be a core dependency #673
- Replace `supported_problem_types` field on pipelines with `problem_type` attribute on base classes #678
- Changed AutoML to only show best results for a given pipeline template in rankings, added `full_rankings` property to show all #682
- Update `ModelFamily` values: don’t list `xgboost`/`catboost` as classifiers now that we have regression pipelines for them #677
- Changed AutoML’s `describe_pipeline` to get problem type from pipeline instead #685
- Standardize `import_or_raise` error messages #683
- Updated argument order of objectives to align with `sklearn`’s #698
- Renamed `pipeline.feature_importance_graph` to `pipeline.graph_feature_importances` #700
- Moved ROC and confusion matrix methods to `evalml.pipelines.plot_utils` #704
- Renamed `MultiClassificationObjective` to `MulticlassClassificationObjective`, to align with pipeline naming scheme #715

- **Documentation Changes**

- Fixed some sphinx warnings #593
- Fixed docstring for `AutoClassificationSearch` with correct command #599
- Limit `readthedocs` formats to pdf, not `htmlzip` and `epub` #594 #600
- Clean up objectives API documentation #605
- Fixed function on Exploring search results page #604
- Update release process doc #567
- `AutoClassificationSearch` and `AutoRegressionSearch` show inherited methods in API reference #651
- Fixed improperly formatted code in breaking changes for changelog #655

- Added configuration to treat Sphinx warnings as errors #660
- Removed separate plotting section for pipelines in API reference #657, #665
- Have leads example notebook load S3 files using https, so we can delete s3fs dev dependency #664
- Categorized components in API reference and added descriptions for each category #663
- Fixed Sphinx warnings about `BalancedAccuracy` objective #669
- Updated API reference to include missing components and clean up pipeline docstrings #689
- Reorganize API ref, and clarify pipeline sub-titles #688
- Add and update preprocessing utils in API reference #687
- Added inheritance diagrams to API reference #695
- Documented which default objective AutoML optimizes for #699
- Create separate install page #701
- Include more utils in API ref, like `import_or_raise` #704
- Add more color to pipeline documentation #705

• Testing Changes

- Matched install commands of `check_latest_dependencies` test and it's GitHub action #578
- Added Github app to auto assign PR author as assignee #477
- Removed unneeded conda installation of xgboost in windows checkin tests #618
- Update graph tests to always use tmpfile dir #649
- Changelog checkin test workaround for release PRs: If 'future release' section is empty of PR refs, pass check #658
- Add changelog checkin test exception for `dep-update` branch #723

Warning: Breaking Changes

- Pipelines will now no longer take an `objective` parameter during instantiation, and will no longer have an `objective` attribute.
- `fit()` and `predict()` now use an optional `objective` parameter, which is only used in binary classification pipelines to fit for a specific objective.
- `score()` will now use a required `objectives` parameter that is used to determine all the objectives to score on. This differs from the previous behavior, where the pipeline's objective was scored on regardless.
- `score()` will now return one dictionary of all objective scores.
- ROC and ConfusionMatrix plot methods via `Auto(*).plot` have been removed by #615 and are replaced by `roc_curve` and `confusion_matrix` in `evalml.pipelines.plot_utils` in #704
- `normalize_confusion_matrix` has been moved to `evalml.pipelines.plot_utils` #704
- Pipelines `_name` field changed to `custom_name`
- Pipelines `supported_problem_types` field is removed because it is no longer necessary #678
- Updated argument order of objectives' `objective_function` to align with sklearn #698

- `pipeline.feature_importance_graph` has been renamed to `pipeline.graph_feature_importances` in #700
- Removed unsupported MSLE objective #704

v0.8.0 Apr. 1, 2020**• Enhancements**

- Add normalization option and information to confusion matrix #484
- Add util function to drop rows with NaN values #487
- Renamed `PipelineBase.name` as `PipelineBase.summary` and redefined `PipelineBase.name` as class property #491
- Added access to parameters in Pipelines with `PipelineBase.parameters` (used to be return of `PipelineBase.describe`) #501
- Added `fill_value` parameter for `SimpleImputer` #509
- Added functionality to override component hyperparameters and made pipelines take hyperparameters from components #516
- Allow `numpy.random.RandomState` for `random_state` parameters #556

• Fixes

- Removed unused dependency `matplotlib`, and move `category_encoders` to test reqs #572

• Changes

- Undo version cap in XGBoost placed in #402 and allowed all released of XGBoost #407
- Support pandas 1.0.0 #486
- Made all references to the logger static #503
- Refactored `model_type` parameter for components and pipelines to `model_family` #507
- Refactored `problem_types` for pipelines and components into `supported_problem_types` #515
- Moved `pipelines/utils.save_pipeline` and `pipelines/utils.load_pipeline` to `PipelineBase.save` and `PipelineBase.load` #526
- Limit number of categories encoded by `OneHotEncoder` #517

• Documentation Changes

- Updated API reference to remove `PipelinePlot` and added moved `PipelineBase` plotting methods #483
- Add code style and github issue guides #463 #512
- Updated API reference for to surface class variables for pipelines and components #537
- Fixed README documentation link #535
- Unhid PR references in changelog #656

• Testing Changes

- Added automated dependency check PR #482, #505
- Updated automated dependency check comment #497

- Have build_docs job use python executor, so that env vars are set properly [#547](#)
- Added simple test to make sure OneHotEncoder's top_n works with large number of categories [#552](#)
- Run windows unit tests on PRs [#557](#)

Warning: Breaking Changes

- AutoClassificationSearch and AutoRegressionSearch's model_types parameter has been refactored into allowed_model_families
- ModelTypes enum has been changed to ModelFamily
- Components and Pipelines now have a model_family field instead of model_type
- get_pipelines utility function now accepts model_families as an argument instead of model_types
- PipelineBase.name no longer returns structure of pipeline and has been replaced by PipelineBase.summary
- PipelineBase.problem_types and Estimator.problem_types has been renamed to supported_problem_types
- pipelines/utils.save_pipeline and pipelines/utils.load_pipeline moved to PipelineBase.save and PipelineBase.load

v0.7.0 Mar. 9, 2020• **Enhancements**

- Added emacs buffers to .gitignore [#350](#)
- Add CatBoost (gradient-boosted trees) classification and regression components and pipelines [#247](#)
- Added Tuner abstract base class [#351](#)
- Added n_jobs as parameter for AutoClassificationSearch and AutoRegressionSearch [#403](#)
- Changed colors of confusion matrix to shades of blue and updated axis order to match scikit-learn's [#426](#)
- Added PipelineBase .graph and .feature_importance_graph methods, moved from previous location [#423](#)
- Added support for python 3.8 [#462](#)

• **Fixes**

- Fixed ROC and confusion matrix plots not being calculated if user passed own additional_objectives [#276](#)
- Fixed ReadtheDocs FileNotFoundError exception for fraud dataset [#439](#)

• **Changes**

- Added n_estimators as a tunable parameter for XGBoost [#307](#)
- Remove unused parameter ObjectiveBase.fit_needs_proba [#320](#)
- Remove extraneous parameter component_type from all components [#361](#)

- Remove unused `rankings.csv` file #397
- Downloaded demo and test datasets so unit tests can run offline #408
- Remove `_needs_fitting` attribute from Components #398
- Changed `plot.feature_importance` to show only non-zero feature importances by default, added optional parameter to show all #413
- Refactored `PipelineBase` to take in parameter dictionary and moved pipeline metadata to class attribute #421
- Dropped support for Python 3.5 #438
- Removed unused `apply.py` file #449
- Clean up `requirements.txt` to remove unused deps #451
- Support installation without all required dependencies #459
- **Documentation Changes**
 - Update `release.md` with instructions to release to internal license key #354
- **Testing Changes**
 - Added tests for utils (and moved current utils to `gen_utils`) #297
 - Moved XGBoost install into it's own separate step on Windows using Conda #313
 - Rewind pandas version to before 1.0.0, to diagnose test failures for that version #325
 - Added dependency update checkin test #324
 - Rewind XGBoost version to before 1.0.0 to diagnose test failures for that version #402
 - Update dependency check to use a whitelist #417
 - Update unit test jobs to not install dev deps #455

Warning: Breaking Changes

- Python 3.5 will not be actively supported.

v0.6.0 Dec. 16, 2019

- **Enhancements**
 - Added ability to create a plot of feature importances #133
 - Add early stopping to AutoML using patience and tolerance parameters #241
 - Added ROC and confusion matrix metrics and plot for classification problems and introduce `PipelineSearchPlots` class #242
 - Enhanced AutoML results with search order #260
 - Added utility function to show system and environment information #300
- **Fixes**
 - Lower botocore requirement #235
 - Fixed `decision_function` calculation for `FraudCost` objective #254
 - Fixed return value of `Recall` metrics #264
 - Components return `self` on fit #289

- **Changes**

- Renamed `automl` classes to `AutoRegressionSearch` and `AutoClassificationSearch` #287
- Updating demo datasets to retain column names #223
- Moving pipeline visualization to `PipelinePlot` class #228
- Standardizing inputs as `pd.DataFrame` / `pd.Series` #130
- Enforcing that pipelines must have an estimator as last component #277
- Added `ipywidgets` as a dependency in `requirements.txt` #278
- Added Random and Grid Search Tuners #240

- **Documentation Changes**

- Adding class properties to API reference #244
- Fix and filter `FutureWarnings` from `scikit-learn` #249, #257
- Adding Linear Regression to API reference and cleaning up some Sphinx warnings #227

- **Testing Changes**

- Added support for testing on Windows with CircleCI #226
- Added support for doctests #233

Warning: Breaking Changes

- The `fit()` method for `AutoClassifier` and `AutoRegressor` has been renamed to `search()`.
- `AutoClassifier` has been renamed to `AutoClassificationSearch`
- `AutoRegressor` has been renamed to `AutoRegressionSearch`
- `AutoClassificationSearch.results` and `AutoRegressionSearch.results` now is a dictionary with `pipeline_results` and `search_order` keys. `pipeline_results` can be used to access a dictionary that is identical to the old `.results` dictionary. Whereas, `search_order` returns a list of the search order in terms of `pipeline_id`.
- Pipelines now require an estimator as the last component in `component_list`. Slicing pipelines now throws an `NotImplementedError` to avoid returning pipelines without an estimator.

v0.5.2 Nov. 18, 2019

- **Enhancements**

- Adding basic pipeline structure visualization #211

- **Documentation Changes**

- Added notebooks to build process #212

v0.5.1 Nov. 15, 2019

- **Enhancements**

- Added basic outlier detection guardrail #151
- Added basic ID column guardrail #135
- Added support for unlimited pipelines with a `max_time` limit #70

- Updated `.readthedocs.yaml` to successfully build [#188](#)
- **Fixes**
 - Removed MSLE from default additional objectives [#203](#)
 - Fixed `random_state` passed in pipelines [#204](#)
 - Fixed slow down in `RFRegressor` [#206](#)
- **Changes**
 - Pulled information for `describe_pipeline` from pipeline's new `describe` method [#190](#)
 - Refactored pipelines [#108](#)
 - Removed guardrails from `Auto(*)` [#202](#), [#208](#)
- **Documentation Changes**
 - Updated documentation to show `max_time` enhancements [#189](#)
 - Updated release instructions for RTD [#193](#)
 - Added notebooks to build process [#212](#)
 - Added contributing instructions [#213](#)
 - Added new content [#222](#)

v0.5.0 Oct. 29, 2019

- **Enhancements**
 - Added basic one hot encoding [#73](#)
 - Use enums for `model_type` [#110](#)
 - Support for splitting regression datasets [#112](#)
 - Auto-infer multiclass classification [#99](#)
 - Added support for other units in `max_time` [#125](#)
 - Detect highly null columns [#121](#)
 - Added additional regression objectives [#100](#)
 - Show an interactive iteration vs. score plot when using `fit()` [#134](#)
- **Fixes**
 - Reordered `describe_pipeline` [#94](#)
 - Added type check for `model_type` [#109](#)
 - Fixed `s` units when setting string `max_time` [#132](#)
 - Fix objectives not appearing in API documentation [#150](#)
- **Changes**
 - Reorganized tests [#93](#)
 - Moved logging to its own module [#119](#)
 - Show progress bar history [#111](#)
 - Using `cloudpickle` instead of `pickle` to allow unloading of custom objectives [#113](#)
 - Removed `render.py` [#154](#)

- **Documentation Changes**

- Update release instructions [#140](#)
- Include additional_objectives parameter [#124](#)
- Added Changelog [#136](#)

- **Testing Changes**

- Code coverage [#90](#)
- Added CircleCI tests for other Python versions [#104](#)
- Added doc notebooks as tests [#139](#)
- Test metadata for CircleCI and 2 core parallelism [#137](#)

v0.4.1 Sep. 16, 2019

- **Enhancements**

- Added AutoML for classification and regressor using Autobase and Skopt [#7](#) [#9](#)
- Implemented standard classification and regression metrics [#7](#)
- Added logistic regression, random forest, and XGBoost pipelines [#7](#)
- Implemented support for custom objectives [#15](#)
- Feature importance for pipelines [#18](#)
- Serialization for pipelines [#19](#)
- Allow fitting on objectives for optimal threshold [#27](#)
- Added detect label leakage [#31](#)
- Implemented callbacks [#42](#)
- Allow for multiclass classification [#21](#)
- Added support for additional objectives [#79](#)

- **Fixes**

- Fixed feature selection in pipelines [#13](#)
- Made random_seed usage consistent [#45](#)

- **Documentation Changes**

- Documentation Changes
- Added docstrings [#6](#)
- Created notebooks for docs [#6](#)
- Initialized readthedocs EvalML [#6](#)
- Added favicon [#38](#)

- **Testing Changes**

- Added testing for loading data [#39](#)

v0.2.0 Aug. 13, 2019

- **Enhancements**

- Created fraud detection objective [#4](#)

v0.1.0 July. 31, 2019

- *First Release*
- **Enhancements**
 - Added lead scoring objective #1
 - Added basic classifier #1
- **Documentation Changes**
 - Initialized Sphinx for docs #1

PYTHON MODULE INDEX

e

`evalml`, 166
`evalml.automl`, 166
`evalml.automl.automl_algorithm`, 166
`evalml.automl.automl_algorithm.automl_algorithm`, 166
`evalml.automl.automl_algorithm.default_algorithm`, 167
`evalml.automl.automl_algorithm.iterative_algorithm`, 169
`evalml.automl.automl_search`, 191
`evalml.automl.callbacks`, 199
`evalml.automl.engine`, 176
`evalml.automl.engine.cf_engine`, 176
`evalml.automl.engine.dask_engine`, 179
`evalml.automl.engine.engine_base`, 181
`evalml.automl.engine.sequential_engine`, 184
`evalml.automl.pipeline_search_plots`, 200
`evalml.automl.utils`, 200
`evalml.data_checks`, 212
`evalml.data_checks.class_imbalance_data_check`, 212
`evalml.data_checks.data_check`, 213
`evalml.data_checks.data_check_action`, 214
`evalml.data_checks.data_check_action_code`, 215
`evalml.data_checks.data_check_message`, 216
`evalml.data_checks.data_check_message_code`, 217
`evalml.data_checks.data_check_message_type`, 219
`evalml.data_checks.data_checks`, 220
`evalml.data_checks.datetime_format_data_check`, 220
`evalml.data_checks.datetime_nan_data_check`, 222
`evalml.data_checks.default_data_checks`, 223
`evalml.data_checks.highly_null_data_check`, 224
`evalml.data_checks.id_columns_data_check`, 226
`evalml.data_checks.invalid_targets_data_check`, 227
`evalml.data_checks.multicollinearity_data_check`, 228
`evalml.data_checks.natural_language_nan_data_check`, 229
`evalml.data_checks.no_variance_data_check`, 231
`evalml.data_checks.outliers_data_check`, 232
`evalml.data_checks.sparsity_data_check`, 233
`evalml.data_checks.target_distribution_data_check`, 235
`evalml.data_checks.target_leakage_data_check`, 236
`evalml.data_checks.uniqueness_data_check`, 237
`evalml.data_checks.utils`, 239
`evalml.demos`, 260
`evalml.demos.breast_cancer`, 260
`evalml.demos.churn`, 261
`evalml.demos.diabetes`, 261
`evalml.demos.fraud`, 262
`evalml.demos.weather`, 262
`evalml.demos.wine`, 263
`evalml.exceptions`, 264
`evalml.exceptions.exceptions`, 264
`evalml.model_family`, 269
`evalml.model_family.model_family`, 269
`evalml.model_family.utils`, 270
`evalml.model_understanding`, 271
`evalml.model_understanding.force_plots`, 278
`evalml.model_understanding.graphs`, 279
`evalml.model_understanding.permutation_importance`, 289
`evalml.model_understanding.prediction_explanations`, 271

[evalml.model_understanding.prediction_explanations.explainers,](#)
[271](#) [evalml.model_understanding.prediction_explanations.explainers,](#)
[evalml.objectives,](#) [303](#) [evalml.objectives,](#) [303](#)
[evalml.objectives.binary_classification_objective,](#) [303](#)
[evalml.objectives.cost_benefit_matrix,](#) [306](#)
[evalml.objectives.fraud_cost,](#) [309](#)
[evalml.objectives.lead_scoring,](#) [312](#)
[evalml.objectives.multiclass_classification_objective,](#) [314](#)
[evalml.objectives.objective_base,](#) [317](#)
[evalml.objectives.regression_objective,](#) [319](#)
[evalml.objectives.sensitivity_low_alert,](#) [321](#)
[evalml.objectives.standard_metrics,](#) [324](#)
[evalml.objectives.time_series_regression_objective,](#) [387](#)
[evalml.objectives.utils,](#) [389](#)
[evalml.pipelines,](#) [474](#)
[evalml.pipelines.binary_classification_pipeline,](#) [1064](#)
[evalml.pipelines.binary_classification_pipeline,](#) [1069](#)
[evalml.pipelines.classification_pipeline,](#) [1070](#)
[evalml.pipelines.component_graph,](#) [1075](#)
[evalml.pipelines.components,](#) [474](#)
[evalml.pipelines.components.component_base,](#) [918](#)
[evalml.pipelines.components.component_base,](#) [920](#)
[evalml.pipelines.components.ensemble,](#) [474](#)
[evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_base,](#) [474](#)
[evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_base,](#) [477](#)
[evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_regressor,](#) [480](#)
[evalml.pipelines.components.estimators,](#) [491](#)
[evalml.pipelines.components.estimators.classifier,](#) [491](#)
[evalml.pipelines.components.estimators.classifier_pipeline_base,](#) [491](#)
[evalml.pipelines.components.estimators.classifier_pipeline_base,](#) [494](#)
[evalml.pipelines.components.estimators.classifier_pipeline_base,](#) [497](#)
[evalml.pipelines.components.estimators.classifier_pipeline_base,](#) [500](#)
[evalml.pipelines.components.estimators.classifier_pipeline_base,](#) [500](#)
[evalml.pipelines.components.estimators.classifiers,](#) [503](#)
[evalml.pipelines.components.estimators.classifiers,](#) [506](#)
[evalml.pipelines.components.estimators.classifiers,](#) [509](#)
[evalml.pipelines.components.estimators.classifiers,](#) [513](#)
[evalml.pipelines.components.estimators.classifiers,](#) [516](#)
[evalml.pipelines.components.estimators.classifiers,](#) [518](#)
[evalml.pipelines.components.estimators.classifiers,](#) [521](#)
[evalml.pipelines.components.estimators.estimator,](#) [628](#)
[evalml.pipelines.components.estimators.regressors,](#) [555](#)
[evalml.pipelines.components.estimators.regressors,](#) [555](#)
[evalml.pipelines.components.estimators.regressors,](#) [559](#)
[evalml.pipelines.components.estimators.regressors,](#) [561](#)
[evalml.pipelines.components.estimators.regressors,](#) [564](#)
[evalml.pipelines.components.estimators.regressors,](#) [568](#)
[evalml.pipelines.components.estimators.regressors,](#) [570](#)
[evalml.pipelines.components.estimators.regressors,](#) [574](#)
[evalml.pipelines.components.estimators.regressors,](#) [577](#)
[evalml.pipelines.components.estimators.regressors,](#) [580](#)
[evalml.pipelines.components.estimators.regressors,](#) [583](#)
[evalml.pipelines.components.estimators.regressors,](#) [585](#)
[evalml.pipelines.components.estimators.regressors,](#) [588](#)
[evalml.pipelines.components.estimators.regressors,](#) [591](#)
[evalml.pipelines.components.transformers,](#) [697](#)
[evalml.pipelines.components.transformers.column_selector,](#) [836](#)
[evalml.pipelines.components.transformers.dimensionality_reduction,](#) [697](#)
[evalml.pipelines.components.transformers.dimensionality_reduction,](#) [697](#)
[evalml.pipelines.components.transformers.dimensionality_reduction,](#) [700](#)
[evalml.pipelines.components.transformers.encoders,](#)

707
 evalml.pipelines.components.transformerselectors.components.transformers.scalers,
 707
 evalml.pipelines.components.transformerselectors.components.transformers.scalers.st
 711
 evalml.pipelines.components.transformerselectors.components.transformers.transformers
 720
 evalml.pipelines.components.transformerselectors.components.transformers.transformers
 720
 evalml.pipelines.components.transformers.feature_selection.rf_classifier_feature_selector,
 723
 evalml.pipelines.components.transformerselectors.components.pipeline_base, 1084
 evalml.pipelines.components.transformerselectors.components.pipeline_regression_feature_selector,
 726
 evalml.pipelines.components.transformers.imputers,
 738
 evalml.pipelines.components.transformers.imputers.imputer,
 738
 evalml.pipelines.components.transformers.imputers.per_column_imputer,
 741
 evalml.pipelines.components.transformers.imputers.simple_imputer,
 744
 evalml.pipelines.components.transformerselectors.components.preprocessing,
 746
 evalml.pipelines.components.transformers.preprocessing,
 760
 evalml.pipelines.components.transformers.preprocessing.datetime_featurizer,
 760
 evalml.pipelines.components.transformers.preprocessing.delayed_feature_transformer,
 763
 evalml.pipelines.components.transformers.preprocessing.drop_null_columns,
 766
 evalml.pipelines.components.transformers.preprocessing.drop_rows_transformer,
 768
 evalml.pipelines.components.transformerselectors.components.problem_types,
 771
 evalml.pipelines.components.transformerselectors.components.problem_types.transformers,
 773
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1278
 776
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1279
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1281
 778
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1283
 781
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1286
 783
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1289
 786
 evalml.pipelines.components.transformerselectors.components.grid_search_tuner, 1294
 818
 evalml.pipelines.components.transformers.samplers.base_sampler,
 818
 evalml.pipelines.components.transformers.samplers.oversampler,
 820
 evalml.pipelines.components.transformers.samplers.undersampler,

A

- `abs_error()` (in module *evalml.pipelines.components.utils*), 921
- `all_components()` (in module *evalml.pipelines.components.utils*), 921
- `all_problem_types()` (in module *evalml.problem_types.problem_types*), 1273
- `AccuracyBinary` (class in *evalml.objectives*), 392
- `AccuracyBinary` (class in *evalml.objectives.standard_metrics*), 325
- `AccuracyMulticlass` (class in *evalml.objectives*), 394
- `AccuracyMulticlass` (class in *evalml.objectives.standard_metrics*), 327
- `add()` (*evalml.tuners.grid_search_tuner.GridSearchTuner* method), 1279
- `add()` (*evalml.tuners.GridSearchTuner* method), 1284
- `add()` (*evalml.tuners.random_search_tuner.RandomSearchTuner* method), 1280
- `add()` (*evalml.tuners.RandomSearchTuner* method), 1285
- `add()` (*evalml.tuners.skopt_tuner.SKOptTuner* method), 1281
- `add()` (*evalml.tuners.SKOptTuner* method), 1286
- `add()` (*evalml.tuners.Tuner* method), 1287
- `add()` (*evalml.tuners.tuner.Tuner* method), 1282
- `add_result()` (*evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm* method), 167
- `add_result()` (*evalml.automl.automl_algorithm.AutoMLAlgorithm* method), 172
- `add_result()` (*evalml.automl.automl_algorithm.default_algorithm.DefaultAlgorithm* method), 169
- `add_result()` (*evalml.automl.automl_algorithm.DefaultAlgorithm* method), 174
- `add_result()` (*evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm* method), 171
- `add_result()` (*evalml.automl.automl_algorithm.IterativeAlgorithm* method), 175
- `add_to_rankings()` (*evalml.automl.automl_search.AutoMLSearch* method), 195
- `add_to_rankings()` (*evalml.automl.AutoMLSearch* method), 206
- `add_to_rankings()` (*evalml.AutoMLSearch* method), 1302
- `all_components()` (in module *evalml.pipelines.components.utils*), 921
- `all_problem_types()` (in module *evalml.problem_types.problem_types*), 1273
- `all_problem_types()` (in module *evalml.problem_types.problem_types* method), 1277
- `allowed_model_families()` (in module *evalml.pipelines.components.utils*), 921
- `ARIMARegressor` (class in *evalml.pipelines*), 1125
- `ARIMARegressor` (class in *evalml.pipelines.components*), 927
- `ARIMARegressor` (class in *evalml.pipelines.components.estimators*), 632
- `ARIMARegressor` (class in *evalml.pipelines.components.estimators.regressors*), 594
- `ARIMARegressor` (class in *evalml.pipelines.components.estimators.regressors.arima_regressors*), 556
- `AUC` (class in *evalml.objectives*), 396
- `AUC` (class in *evalml.objectives.standard_metrics*), 329
- `AUCMacro` (class in *evalml.objectives*), 398
- `AUCMacro` (class in *evalml.objectives.standard_metrics*), 331
- `AUCMicro` (class in *evalml.objectives*), 400
- `AUCMicro` (class in *evalml.objectives.standard_metrics*), 333
- `AUCWeighted` (class in *evalml.objectives*), 401
- `AUCWeighted` (class in *evalml.objectives.standard_metrics*), 334
- `AutoMLAlgorithm` (class in *evalml.automl.automl_algorithm*), 172
- `AutoMLAlgorithm` (class in *evalml.automl.automl_algorithm.automl_algorithm*), 166
- `AutoMLAlgorithmException`, 167, 173
- `AutoMLConfig` (in module *evalml.automl.utils*), 201
- `AutoMLSearch` (class in *evalml*), 1299
- `AutoMLSearch` (class in *evalml.automl*), 203

AutoMLSearch (class
evalml.automl.automl_search), 192
 AutoMLSearchException, 265, 267

B

BalancedAccuracyBinary (class
evalml.objectives), 403
 BalancedAccuracyBinary (class
evalml.objectives.standard_metrics), 336
 BalancedAccuracyMulticlass (class
evalml.objectives), 405
 BalancedAccuracyMulticlass (class
evalml.objectives.standard_metrics), 338
 BalancedClassificationSampler (class
evalml.preprocessing.data_splitters), 1266
 BalancedClassificationSampler (class
evalml.preprocessing.data_splitters.balanced_classification_sampler),
 1262
 BaselineClassifier (class
evalml.pipelines.components), 929
 BaselineClassifier (class
evalml.pipelines.components.estimators),
 634
 BaselineClassifier (class
evalml.pipelines.components.estimators.classifiers),
 524
 BaselineClassifier (class
evalml.pipelines.components.estimators.classifiers.baseline_classifier),
 491
 BaselineRegressor (class
evalml.pipelines.components), 932
 BaselineRegressor (class
evalml.pipelines.components.estimators),
 637
 BaselineRegressor (class
evalml.pipelines.components.estimators.regressors),
 597
 BaselineRegressor (class
evalml.pipelines.components.estimators.regressors.baseline_regressor),
 559
 BaseMeta (class in *evalml.utils.base_meta*), 1288
 BaseSampler (class
evalml.pipelines.components.transformers.samplers.base_sampler),
 818

in *batch_number()* (*evalml.automl.automl_algorithm.IterativeAlgorithm*
property), 175
best_pipeline() (*evalml.automl.automl_search.AutoMLSearch*
property), 195
best_pipeline() (*evalml.automl.AutoMLSearch*
property), 206
best_pipeline() (*evalml.AutoMLSearch* *property*),
 1302
binary_objective_vs_threshold() (in *mod-*
ule evalml.model_understanding), 292
binary_objective_vs_threshold() (in *mod-*
ule evalml.model_understanding.graphs), 280
 BinaryClassificationObjective (class in
evalml.objectives), 407
 BinaryClassificationObjective (class in
evalml.objectives.binary_classification_objective),
 303
 BinaryClassificationPipeline (class in
evalml.pipelines), 1128
 BinaryClassificationPipeline (class in
evalml.pipelines.binary_classification_pipeline),
 1064
 BinaryClassificationPipelineMixin (class
 in *evalml.pipelines.binary_classification_pipeline_mixin*),
 1070
build_engine_from_str() (in *module*
evalml.automl.automl_search), 197
build_prophet_df()
 (*evalml.pipelines.components.estimators.ProphetRegressor*
static method), 679
build_prophet_df()
 (*evalml.pipelines.components.estimators.regressors.prophet_regr*
static method), 581
build_prophet_df()
 (*evalml.pipelines.components.estimators.regressors.ProphetRegr*
static method), 617
build_prophet_df()
 (*evalml.pipelines.components.ProphetRegressor*
static method), 1013
build_prophet_df()
 (*evalml.pipelines.ProphetRegressor* *static*
method), 1200

C

batch_number() (*evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm*
property), 167
batch_number() (*evalml.automl.automl_algorithm.AutoMLAlgorithm*
method), 393
batch_number() (*evalml.automl.automl_algorithm.default_algorithm.DefaultAlgorithm*
property), 169
batch_number() (*evalml.automl.automl_algorithm.DefaultAlgorithm*
method), 395
batch_number() (*evalml.automl.automl_algorithm.DefaultAlgorithm*
method), 395
batch_number() (*evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm*
property), 171
batch_number() (*evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm*
method), 393
calculate_percent_difference()
 (*evalml.objectives.AccuracyBinary* *class*
method), 393
calculate_percent_difference()
 (*evalml.objectives.AccuracyMulticlass* *class*
method), 395
calculate_percent_difference()
 (*evalml.objectives.AUC* *class method*), 397
calculate_percent_difference()
 (*evalml.objectives.AUCMacro* *class method*),
 397

[399](#)
`calculate_percent_difference()`
 (`evalml.objectives.AUCMicro` class method), [400](#)
`calculate_percent_difference()`
 (`evalml.objectives.AUCWeighted` class method), [402](#)
`calculate_percent_difference()`
 (`evalml.objectives.BalancedAccuracyBinary` class method), [404](#)
`calculate_percent_difference()`
 (`evalml.objectives.BalancedAccuracyMulticlass` class method), [406](#)
`calculate_percent_difference()`
 (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` class method), [304](#)
`calculate_percent_difference()`
 (`evalml.objectives.BinaryClassificationObjective` class method), [408](#)
`calculate_percent_difference()`
 (`evalml.objectives.cost_benefit_matrix.CostBenefitMatrix` class method), [307](#)
`calculate_percent_difference()`
 (`evalml.objectives.CostBenefitMatrix` class method), [410](#)
`calculate_percent_difference()`
 (`evalml.objectives.ExpVariance` class method), [413](#)
`calculate_percent_difference()`
 (`evalml.objectives.F1` class method), [414](#)
`calculate_percent_difference()`
 (`evalml.objectives.F1Macro` class method), [416](#)
`calculate_percent_difference()`
 (`evalml.objectives.F1Micro` class method), [418](#)
`calculate_percent_difference()`
 (`evalml.objectives.F1Weighted` class method), [420](#)
`calculate_percent_difference()`
 (`evalml.objectives.fraud_cost.FraudCost` class method), [310](#)
`calculate_percent_difference()`
 (`evalml.objectives.FraudCost` class method), [422](#)
`calculate_percent_difference()`
 (`evalml.objectives.Gini` class method), [425](#)
`calculate_percent_difference()`
 (`evalml.objectives.lead_scoring.LeadScoring` class method), [312](#)
`calculate_percent_difference()`
 (`evalml.objectives.LeadScoring` class method), [427](#)
`calculate_percent_difference()`
 (`evalml.objectives.LogLossBinary` class method), [429](#)
`calculate_percent_difference()`
 (`evalml.objectives.LogLossMulticlass` class method), [431](#)
`calculate_percent_difference()`
 (`evalml.objectives.MAE` class method), [433](#)
`calculate_percent_difference()`
 (`evalml.objectives.MAPE` class method), [435](#)
`calculate_percent_difference()`
 (`evalml.objectives.MaxError` class method), [436](#)
`calculate_percent_difference()`
 (`evalml.objectives.MulticlassClassificationObjective` class method), [438](#)
`calculate_percent_difference()`
 (`evalml.objectives.MCCBinary` class method), [440](#)
`calculate_percent_difference()`
 (`evalml.objectives.MCCMulticlass` class method), [440](#)
`calculate_percent_difference()`
 (`evalml.objectives.MeanSquaredLogError` class method), [442](#)
`calculate_percent_difference()`
 (`evalml.objectives.MedianAE` class method), [444](#)
`calculate_percent_difference()`
 (`evalml.objectives.MSE` class method), [445](#)
`calculate_percent_difference()`
 (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` class method), [315](#)
`calculate_percent_difference()`
 (`evalml.objectives.MulticlassClassificationObjective` class method), [447](#)
`calculate_percent_difference()`
 (`evalml.objectives.objective_base.ObjectiveBase` class method), [317](#)
`calculate_percent_difference()`
 (`evalml.objectives.ObjectiveBase` class method), [449](#)
`calculate_percent_difference()`
 (`evalml.objectives.Precision` class method), [451](#)
`calculate_percent_difference()`
 (`evalml.objectives.PrecisionMacro` class method), [453](#)
`calculate_percent_difference()`
 (`evalml.objectives.PrecisionMicro` class method), [455](#)
`calculate_percent_difference()`
 (`evalml.objectives.PrecisionWeighted` class method), [457](#)
`calculate_percent_difference()`
 (`evalml.objectives.R2` class method), [458](#)
`calculate_percent_difference()`

<code>(evalml.objectives.Recall class method), 460</code>	<code>(evalml.objectives.standard_metrics.ExpVariance class method), 340</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RecallMacro class method), 462</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.F1 class method), 342</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RecallMicro class method), 464</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.F1Macro class method), 344</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RecallWeighted class method), 465</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.F1Micro class method), 346</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.regression_objective.RegressionObjective class method), 320</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.F1Weighted class method), 347</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RegressionObjective class method), 467</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.Gini class method), 349</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RootMeanSquaredError class method), 469</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.LogLossBinary class method), 351</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.RootMeanSquaredLogError class method), 471</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.LogLossMulticlass class method), 353</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.sensitivity_low_alert.SensitivityLowAlert class method), 322</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MAE class method), 355</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.SensitivityLowAlert class method), 472</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MAPE class method), 357</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AccuracyBinary class method), 326</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MaxError class method), 358</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AccuracyMulticlass class method), 328</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MCCBinary class method), 360</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AUC class method), 329</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MCCMulticlass class method), 362</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AUCMacro class method), 332</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MeanSquaredLogError class method), 364</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AUCMicro class method), 333</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MedianAE class method), 365</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.AUCWeighted class method), 335</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.MSE class method), 367</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.BalancedAccuracyBinary class method), 336</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.Precision class method), 368</code>
<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.BalancedAccuracyMulticlass class method), 339</code>	<code>calculate_percent_difference()</code> <code>(evalml.objectives.standard_metrics.PrecisionMacro class method), 371</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>

(*evalml.objectives.standard_metrics.PrecisionMicro* class method), 372

(*evalml.objectives.cost_benefit_matrix.CostBenefitMatrix* property), 307

calculate_percent_difference() (*evalml.objectives.standard_metrics.PrecisionWeighted* class method), 374

calculate_percent_difference() (*evalml.objectives.standard_metrics.R2* class method), 375

calculate_percent_difference() (*evalml.objectives.standard_metrics.Recall* class method), 377

calculate_percent_difference() (*evalml.objectives.standard_metrics.RecallMacro* class method), 379

calculate_percent_difference() (*evalml.objectives.standard_metrics.RecallMicro* class method), 381

calculate_percent_difference() (*evalml.objectives.standard_metrics.RecallWeighted* class method), 383

calculate_percent_difference() (*evalml.objectives.standard_metrics.RootMeanSquaredError* class method), 384

calculate_percent_difference() (*evalml.objectives.standard_metrics.RootMeanSquaredLogError* class method), 386

calculate_percent_difference() (*evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective* class method), 388

calculate_permutation_importance() (in module *evalml.model_understanding*), 292

calculate_permutation_importance() (in module *evalml.model_understanding.permutation_importance*), 290

calculate_permutation_importance_one_column() (in module *evalml.model_understanding*), 293

calculate_permutation_importance_one_column() (in module *evalml.model_understanding.permutation_importance*), 290

can_optimize_threshold() (*evalml.objectives.AccuracyBinary* property), 393

can_optimize_threshold() (*evalml.objectives.AUC* property), 397

can_optimize_threshold() (*evalml.objectives.BalancedAccuracyBinary* property), 404

can_optimize_threshold() (*evalml.objectives.binary_classification_objective.BinaryClassificationObjective* property), 304

can_optimize_threshold() (*evalml.objectives.BinaryClassificationObjective* property), 408

can_optimize_threshold()

can_optimize_threshold() (*evalml.objectives.cost_benefit_matrix.CostBenefitMatrix* property), 307

can_optimize_threshold() (*evalml.objectives.CostBenefitMatrix* property), 411

can_optimize_threshold() (*evalml.objectives.F1* property), 415

can_optimize_threshold() (*evalml.objectives.fraud_cost.FraudCost* property), 310

can_optimize_threshold() (*evalml.objectives.FraudCost* property), 422

can_optimize_threshold() (*evalml.objectives.Gini* property), 425

can_optimize_threshold() (*evalml.objectives.lead_scoring.LeadScoring* property), 313

can_optimize_threshold() (*evalml.objectives.LeadScoring* property), 427

can_optimize_threshold() (*evalml.objectives.LogLossBinary* property), 430

can_optimize_threshold() (*evalml.objectives.MCCBinary* property), 438

can_optimize_threshold() (*evalml.objectives.Precision* property), 451

can_optimize_threshold() (*evalml.objectives.Recall* property), 460

can_optimize_threshold() (*evalml.objectives.sensitivity_low_alert.SensitivityLowAlert* property), 323

can_optimize_threshold() (*evalml.objectives.SensitivityLowAlert* property), 473

can_optimize_threshold() (*evalml.objectives.standard_metrics.AccuracyBinary* property), 326

can_optimize_threshold() (*evalml.objectives.standard_metrics.AUC* property), 330

can_optimize_threshold() (*evalml.objectives.standard_metrics.BalancedAccuracyBinary* property), 337

can_optimize_threshold() (*evalml.objectives.standard_metrics.F1* property), 415

can_optimize_threshold() (*evalml.objectives.standard_metrics.Gini* property), 349

can_optimize_threshold() (*evalml.objectives.standard_metrics.LogLossBinary*

`property)`, 351
`can_optimize_threshold()`
 (`evalml.objectives.standard_metrics.MCCBinary`
 `property)`, 360
`can_optimize_threshold()`
 (`evalml.objectives.standard_metrics.Precision`
 `property)`, 369
`can_optimize_threshold()`
 (`evalml.objectives.standard_metrics.Recall`
 `property)`, 377
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline`
 `method)`, 1065
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.BinaryClassificationPipeline`
 `method)`, 1129
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.classification_pipeline.ClassificationPipeline`
 `method)`, 1072
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.ClassificationPipeline`
 `method)`, 1139
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline`
 `method)`, 1080
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.MulticlassClassificationPipeline`
 `method)`, 1186
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.pipeline_base.PipelineBase`
 `method)`, 1085
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.PipelineBase` `method)`, 1196
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.regression_pipeline.RegressionPipeline`
 `method)`, 1091
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.RegressionPipeline` `method)`,
 1207
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline`
 `method)`, 1096
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline`
 `method)`, 1102
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline`
 `method)`, 1108
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase`
 `method)`, 1113
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline`
 `method)`, 1119
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.TimeSeriesRegressionPipeline`
 `method)`, 1234
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.TimeSeriesClassificationPipeline`
 `method)`, 1239
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline`
 `method)`, 1245
`can_tune_threshold_with_objective()`
 (`evalml.pipelines.TimeSeriesRegressionPipeline`
 `method)`, 1250
`cancel()` (`evalml.automl.engine.cf_engine.CFComputation`
 `method)`, 177
`cancel()` (`evalml.automl.engine.dask_engine.DaskComputation`
 `method)`, 179
`cancel()` (`evalml.automl.engine.engine_base.EngineComputation`
 `method)`, 182
`cancel()` (`evalml.automl.engine.EngineComputation`
 `method)`, 189
`cancel()` (`evalml.automl.engine.sequential_engine.SequentialComputation`
 `method)`, 185
`CatBoostClassifier` (class in `evalml.pipelines`),
 in `evalml.pipelines.components`, 934
`CatBoostClassifier` (class in
 `evalml.pipelines.components.estimators`),
 639
`CatBoostClassifier` (class in
 `evalml.pipelines.components.estimators.classifiers`),
 527
`CatBoostClassifier` (class in
 `evalml.pipelines.components.estimators.classifiers.catboost_class`
 494
`CatBoostRegressor` (class in `evalml.pipelines`),
 1135
`CatBoostRegressor` (class in
 `evalml.pipelines.components`), 937
`CatBoostRegressor` (class in
 `evalml.pipelines.components.estimators`),
 642
`CatBoostRegressor` (class in
 `evalml.pipelines.components.estimators.regressors`),
 599
`CatBoostRegressor` (class in
 `evalml.pipelines.components.estimators.regressors.catboost_regre`
 562
`categories()` (`evalml.pipelines.components.OneHotEncoder`
 `method)`, 1001
`categories()` (`evalml.pipelines.components.transformers.encoders.one`
 `method)`, 709
`categories()` (`evalml.pipelines.components.transformers.encoders.One`
 `method)`, 716

[categories\(\) \(evalml.pipelines.components.transformers.OneHotEncoder method\), 879](#)
[categories\(\) \(evalml.pipelines.OneHotEncoder method\), 1191](#)
[CFClient \(class in evalml.automl.engine.cf_engine\), 176](#)
[CFComputation \(class in evalml.automl.engine.cf_engine\), 176](#)
[CFEngine \(class in evalml.automl.engine\), 186](#)
[CFEngine \(class in evalml.automl.engine.cf_engine\), 177](#)
[check_all_pipeline_names_unique\(\) \(in module evalml.automl.utils\), 201](#)
[check_for_fit\(\) \(evalml.pipelines.components.component_base.BaseComponentBase class method\), 920](#)
[check_for_fit\(\) \(evalml.pipelines.components.ComponentBase class method\), 941](#)
[check_for_fit\(\) \(evalml.pipelines.components.transformers.BaseTransformer class method\), 711](#)
[check_for_fit\(\) \(evalml.pipelines.components.transformers.BaseTransformer class method\), 749](#)
[check_for_fit\(\) \(evalml.pipelines.pipeline_meta.PipelineBase class method\), 1089](#)
[classes_\(\) \(evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline property\), 1065](#)
[classes_\(\) \(evalml.pipelines.BinaryClassificationPipeline property\), 1129](#)
[classes_\(\) \(evalml.pipelines.classification_pipeline.ClassificationPipeline property\), 1072](#)
[classes_\(\) \(evalml.pipelines.ClassificationPipeline property\), 1139](#)
[classes_\(\) \(evalml.pipelines.components.BaselineClassifier property\), 930](#)
[classes_\(\) \(evalml.pipelines.components.BaselineClassifier property\), 930](#)
[classes_\(\) \(evalml.pipelines.components.estimators.BaselineClassifier property\), 635](#)
[classes_\(\) \(evalml.pipelines.components.estimators.classifiers.BaselineClassifier property\), 492](#)
[classes_\(\) \(evalml.pipelines.components.estimators.classifiers.BaselineClassifier property\), 525](#)
[classes_\(\) \(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline property\), 1080](#)
[classes_\(\) \(evalml.pipelines.MulticlassClassificationPipeline property\), 1186](#)
[classes_\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1096](#)
[classes_\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1102](#)
[classes_\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1108](#)
[classes_\(\) \(evalml.pipelines.TimeSeriesBinaryClassificationPipeline property\), 1234](#)
[classes_\(\) \(evalml.pipelines.TimeSeriesClassificationPipeline property\), 1240](#)
[classes_\(\) \(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property\), 1240](#)
[ClassificationPipeline \(class in evalml.pipelines\), 1138](#)
[ClassificationPipeline \(class in evalml.pipelines.classification_pipeline\), 1070](#)
[ClassImbalanceDataCheck \(class in evalml.data_checks\), 241](#)
[ClassImbalanceDataCheck \(class in evalml.data_checks.class_imbalance_data_check\), 212](#)
[classproperty \(class in evalml.utils\), 1296](#)
[classproperty \(class in evalml.utils.gen_utils\), 1290](#)
[clone\(\) \(evalml.pipelines.BinaryClassificationPipeline method\), 1126](#)
[clone\(\) \(evalml.pipelines.BinaryClassificationPipeline method\), 1065](#)
[clone\(\) \(evalml.pipelines.BinaryClassificationPipeline method\), 1129](#)
[clone\(\) \(evalml.pipelines.BinaryClassificationPipeline method\), 1134](#)
[clone\(\) \(evalml.pipelines.CatBoostRegressor method\), 1137](#)
[clone\(\) \(evalml.pipelines.ClassificationPipeline method\), 1072](#)
[clone\(\) \(evalml.pipelines.ClassificationPipeline method\), 1139](#)
[clone\(\) \(evalml.pipelines.components.ARIMARegressor method\), 928](#)
[clone\(\) \(evalml.pipelines.components.BaselineClassifier method\), 930](#)
[clone\(\) \(evalml.pipelines.components.BaselineRegressor method\), 933](#)
[clone\(\) \(evalml.pipelines.components.CatBoostClassifier method\), 935](#)
[clone\(\) \(evalml.pipelines.components.CatBoostRegressor method\), 938](#)
[clone\(\) \(evalml.pipelines.components.component_base.ComponentBase method\), 918](#)
[clone\(\) \(evalml.pipelines.ComponentBase method\), 940](#)
[clone\(\) \(evalml.pipelines.DateTimeFeaturizer method\), 942](#)
[clone\(\) \(evalml.pipelines.DecisionTreeClassifier method\), 945](#)
[clone\(\) \(evalml.pipelines.DecisionTreeRegressor method\), 948](#)
[clone\(\) \(evalml.pipelines.DropNullColumns method\), 951](#)
[clone\(\) \(evalml.pipelines.DFSTransformer method\), 953](#)
[clone\(\) \(evalml.pipelines.DropColumns method\), 955](#)
[clone\(\) \(evalml.pipelines.DropNullColumns method\), 955](#)

method), 957

clone() (evalml.pipelines.components.DropRowsTransformer.method), 959

clone() (evalml.pipelines.components.ElasticNetClassifier.method), 962

clone() (evalml.pipelines.components.ElasticNetRegressor.method), 964

clone() (evalml.pipelines.components.EmailFeaturizer.method), 966

clone() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifier.method), 475

clone() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifierRandomForest.method), 478

clone() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleRegressor.method), 481

clone() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier.method), 484

clone() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier.method), 487

clone() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor.method), 489

clone() (evalml.pipelines.components.Estimator.method), 968

clone() (evalml.pipelines.components.estimators.ARIMARegressor.method), 633

clone() (evalml.pipelines.components.estimators.BaselineClassifier.method), 635

clone() (evalml.pipelines.components.estimators.BaselineRegressor.method), 638

clone() (evalml.pipelines.components.estimators.CatBoostClassifier.method), 640

clone() (evalml.pipelines.components.estimators.CatBoostRegressor.method), 643

clone() (evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier.method), 492

clone() (evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier.method), 525

clone() (evalml.pipelines.components.estimators.classifiers.CatBoostClassifier.method), 495

clone() (evalml.pipelines.components.estimators.classifiers.CatBoostClassifier.method), 528

clone() (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier.method), 498

clone() (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier.method), 531

clone() (evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.ElasticNetClassifier.method), 501

clone() (evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier.method), 534

clone() (evalml.pipelines.components.estimators.classifiers.elasticnet_regressor.ElasticNetRegressor.method), 505

clone() (evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier.method), 537

clone() (evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier.method), 508

clone() (evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier.method), 540

clone() (evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier.method), 511

clone() (evalml.pipelines.components.estimators.classifiers.LightGBMClassifier.method), 543

clone() (evalml.pipelines.components.estimators.classifiers.logistic_regression.LogisticRegressionClassifier.method), 514

clone() (evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier.method), 546

clone() (evalml.pipelines.components.estimators.classifiers.RandomForestClassifier.method), 548

clone() (evalml.pipelines.components.estimators.classifiers.RandomForestRegressor.method), 517

clone() (evalml.pipelines.components.estimators.classifiers.svm_classifier.SVMClassifier.method), 520

clone() (evalml.pipelines.components.estimators.classifiers.SVMClassifier.method), 551

clone() (evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier.method), 522

clone() (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier.method), 554

clone() (evalml.pipelines.components.estimators.DecisionTreeRegressor.method), 646

clone() (evalml.pipelines.components.estimators.DecisionTreeRegressor.method), 649

clone() (evalml.pipelines.components.estimators.ElasticNetClassifier.method), 652

clone() (evalml.pipelines.components.estimators.ElasticNetRegressor.method), 655

clone() (evalml.pipelines.components.estimators.Estimator.method), 657

clone() (evalml.pipelines.components.estimators.estimator.Estimator.method), 629

clone() (evalml.pipelines.components.estimators.ExtraTreesClassifier.method), 660

clone() (evalml.pipelines.components.estimators.ExtraTreesRegressor.method), 663

clone() (evalml.pipelines.components.estimators.KNeighborsClassifier.method), 665

clone() (evalml.pipelines.components.estimators.LightGBMClassifier.method), 668

clone() (evalml.pipelines.components.estimators.LightGBMRegressor.method), 671

clone() (evalml.pipelines.components.estimators.LinearRegressor.method), 674

clone() (evalml.pipelines.components.estimators.LogisticRegressionClassifier.method), 676

clone() (evalml.pipelines.components.estimators.ProphetRegressor.method), 679

clone() (evalml.pipelines.components.estimators.RandomForestClassifier.method), 681

clone() (evalml.pipelines.components.estimators.RandomForestRegressor.method), 681

method), 683

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 686

method), 557

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 688

method), 595

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 690

method), 560

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 693

method), 598

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 696

method), 563

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 600

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 566

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 603

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 569

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 606

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 572

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 609

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 575

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 612

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 578

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 614

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 581

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 617

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 619

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 584

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 587

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 622

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 589

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 624

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 592

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

method), 626

clone () (evalml.pipelines.components.estimators.regressor_ensemble), 697

clone () (evalml.pipelines.components.estimators.SVMClassifier), 1023

`method`), 1025
`clone()` (`evalml.pipelines.components.SelectColumns` `method`), 1027
`clone()` (`evalml.pipelines.components.SimpleImputer` `method`), 1029
`clone()` (`evalml.pipelines.components.SklearnStackedEnsembleClassifier` `method`), 1032
`clone()` (`evalml.pipelines.components.SklearnStackedEnsembleRegressor` `method`), 1035
`clone()` (`evalml.pipelines.components.StandardScaler` `method`), 1037
`clone()` (`evalml.pipelines.components.SVMClassifier` `method`), 1039
`clone()` (`evalml.pipelines.components.SVMRegressor` `method`), 1041
`clone()` (`evalml.pipelines.components.TargetEncoder` `method`), 1044
`clone()` (`evalml.pipelines.components.TargetImputer` `method`), 1046
`clone()` (`evalml.pipelines.components.TextFeaturizer` `method`), 1048
`clone()` (`evalml.pipelines.components.TimeSeriesBaselineEstimator` `method`), 1050
`clone()` (`evalml.pipelines.components.Transformer` `method`), 1053
`clone()` (`evalml.pipelines.components.transformers.column_selector.ColumnSelector` `method`), 837
`clone()` (`evalml.pipelines.components.transformers.column_selector.DuplicateColumns` `method`), 839
`clone()` (`evalml.pipelines.components.transformers.column_selector.SelectedByType` `method`), 841
`clone()` (`evalml.pipelines.components.transformers.column_selector.SelectedColumns` `method`), 843
`clone()` (`evalml.pipelines.components.transformers.DateTimeFeaturizer` `method`), 852
`clone()` (`evalml.pipelines.components.transformers.DelayedFeatureTransformer` `method`), 854
`clone()` (`evalml.pipelines.components.transformers.DFSTransformer` `method`), 857
`clone()` (`evalml.pipelines.components.transformers.dimensionality_reduction.LinearDiscriminantAnalysis` `method`), 698
`clone()` (`evalml.pipelines.components.transformers.dimensionality_reduction.LinearDiscriminantAnalysis` `method`), 703
`clone()` (`evalml.pipelines.components.transformers.dimensionality_reduction.PrincipalComponentAnalysis` `method`), 705
`clone()` (`evalml.pipelines.components.transformers.dimensionality_reduction.PrincipalComponentAnalysis` `method`), 701
`clone()` (`evalml.pipelines.components.transformers.DropColumns` `method`), 859
`clone()` (`evalml.pipelines.components.transformers.DropSubColumns` `method`), 861
`clone()` (`evalml.pipelines.components.transformers.DropRowsFromFeature` `method`), 863
`clone()` (`evalml.pipelines.components.transformers.EmailFeaturizer` `method`), 865
`clone()` (`evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder` `method`), 709
`clone()` (`evalml.pipelines.components.transformers.encoders.OneHotEncoder` `method`), 716
`clone()` (`evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder` `method`), 713
`clone()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder` `method`), 719
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector.FeatureSelector` `method`), 721
`clone()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 731
`clone()` (`evalml.pipelines.components.transformers.feature_selection.rf_classifier.RandomForestClassifier` `method`), 724
`clone()` (`evalml.pipelines.components.transformers.feature_selection.rf_classifier.RandomForestClassifier` `method`), 728
`clone()` (`evalml.pipelines.components.transformers.feature_selection.RFClassifier` `method`), 733
`clone()` (`evalml.pipelines.components.transformers.feature_selection.RFClassifier` `method`), 736
`clone()` (`evalml.pipelines.components.transformers.FeatureSelector` `method`), 867
`clone()` (`evalml.pipelines.components.transformers.Imputer` `method`), 870
`clone()` (`evalml.pipelines.components.transformers.imputers.Imputer` `method`), 751
`clone()` (`evalml.pipelines.components.transformers.imputers.imputer.Imputer` `method`), 739
`clone()` (`evalml.pipelines.components.transformers.imputers.per_column.PerColumnImputer` `method`), 742
`clone()` (`evalml.pipelines.components.transformers.imputers.PerColumnImputer` `method`), 753
`clone()` (`evalml.pipelines.components.transformers.imputers.simple_imputer.SimpleImputer` `method`), 745
`clone()` (`evalml.pipelines.components.transformers.imputers.SimpleImputer` `method`), 756
`clone()` (`evalml.pipelines.components.transformers.imputers.target_imputer.TargetImputer` `method`), 747
`clone()` (`evalml.pipelines.components.transformers.imputers.TargetImputer` `method`), 758
`clone()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis` `method`), 872
`clone()` (`evalml.pipelines.components.transformers.LogTransformer` `method`), 874
`clone()` (`evalml.pipelines.components.transformers.LSA` `method`), 876
`clone()` (`evalml.pipelines.components.transformers.OneHotEncoder` `method`), 879
`clone()` (`evalml.pipelines.components.transformers.Oversampler` `method`), 882
`clone()` (`evalml.pipelines.components.transformers.PCA` `method`), 884
`clone()` (`evalml.pipelines.components.transformers.PerColumnImputer` `method`), 884

[1155](#)
`clone()` (*evalml.pipelines.ElasticNetClassifier method*), [1157](#)
`clone()` (*evalml.pipelines.ElasticNetRegressor method*), [1160](#)
`clone()` (*evalml.pipelines.Estimator method*), [1162](#)
`clone()` (*evalml.pipelines.ExtraTreesClassifier method*), [1165](#)
`clone()` (*evalml.pipelines.ExtraTreesRegressor method*), [1168](#)
`clone()` (*evalml.pipelines.FeatureSelector method*), [1170](#)
`clone()` (*evalml.pipelines.KNeighborsClassifier method*), [1173](#)
`clone()` (*evalml.pipelines.LightGBMClassifier method*), [1176](#)
`clone()` (*evalml.pipelines.LightGBMRegressor method*), [1178](#)
`clone()` (*evalml.pipelines.LinearRegressor method*), [1181](#)
`clone()` (*evalml.pipelines.LogisticRegressionClassifier method*), [1183](#)
`clone()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method*), [1080](#)
`clone()` (*evalml.pipelines.MulticlassClassificationPipeline method*), [1186](#)
`clone()` (*evalml.pipelines.OneHotEncoder method*), [1191](#)
`clone()` (*evalml.pipelines.PerColumnImputer method*), [1193](#)
`clone()` (*evalml.pipelines.pipeline_base.PipelineBase method*), [1085](#)
`clone()` (*evalml.pipelines.PipelineBase method*), [1196](#)
`clone()` (*evalml.pipelines.ProphetRegressor method*), [1200](#)
`clone()` (*evalml.pipelines.RandomForestClassifier method*), [1202](#)
`clone()` (*evalml.pipelines.RandomForestRegressor method*), [1205](#)
`clone()` (*evalml.pipelines.regression_pipeline.RegressionPipeline method*), [1091](#)
`clone()` (*evalml.pipelines.RegressionPipeline method*), [1207](#)
`clone()` (*evalml.pipelines.RFClassifierSelectFromModel method*), [1211](#)
`clone()` (*evalml.pipelines.RFRegressorSelectFromModel method*), [1214](#)
`clone()` (*evalml.pipelines.SimpleImputer method*), [1216](#)
`clone()` (*evalml.pipelines.SklearnStackedEnsembleClassifier method*), [1219](#)
`clone()` (*evalml.pipelines.SklearnStackedEnsembleRegressor method*), [1222](#)
`clone()` (*evalml.pipelines.StandardScaler method*), [1224](#)
`clone()` (*evalml.pipelines.SVMClassifier method*), [1226](#)
`clone()` (*evalml.pipelines.SVMRegressor method*), [1228](#)
`clone()` (*evalml.pipelines.TargetEncoder method*), [1231](#)
`clone()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), [1096](#)
`clone()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), [1102](#)
`clone()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), [1108](#)
`clone()` (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase method*), [1113](#)
`clone()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method*), [1119](#)
`clone()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline method*), [1234](#)
`clone()` (*evalml.pipelines.TimeSeriesClassificationPipeline method*), [1240](#)
`clone()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method*), [1245](#)
`clone()` (*evalml.pipelines.TimeSeriesRegressionPipeline method*), [1251](#)
`clone()` (*evalml.pipelines.Transformer method*), [1255](#)
`clone()` (*evalml.pipelines.XGBoostClassifier method*), [1257](#)
`clone()` (*evalml.pipelines.XGBoostRegressor method*), [1260](#)
`close()` (*evalml.automl.engine.cf_engine.CFClient method*), [176](#)
`close()` (*evalml.automl.engine.cf_engine.CFEngine method*), [178](#)
`close()` (*evalml.automl.engine.CFEngine method*), [186](#)
`close()` (*evalml.automl.engine.dask_engine.DaskEngine method*), [180](#)
`close()` (*evalml.automl.engine.DaskEngine method*), [188](#)
`close()` (*evalml.automl.engine.sequential_engine.SequentialEngine method*), [185](#)
`close()` (*evalml.automl.engine.SequentialEngine method*), [190](#)
`close()` (*evalml.automl.SequentialEngine method*), [211](#)
`close_engine()` (*evalml.automl.automl_search.AutoMLSearch method*), [195](#)
`close_engine()` (*evalml.automl.AutoMLSearch method*), [206](#)
`close_engine()` (*evalml.AutoMLSearch method*), [1302](#)
`ColumnSelector` (class in *evalml.pipelines.components.transformers.column_selectors*), [1370](#)

836

ComponentBase (class in *evalml.pipelines.components*), 939

ComponentBase (class in *evalml.pipelines.components.component_base*), 918

ComponentBaseMeta (class in *evalml.pipelines.components*), 941

ComponentBaseMeta (class in *evalml.pipelines.components.component_base_meta*), 920

ComponentGraph (class in *evalml.pipelines*), 1143

ComponentGraph (class in *evalml.pipelines.component_graph*), 1076

ComponentNotYetFittedError, 265, 267

compute_estimator_features() (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1066

compute_estimator_features() (*evalml.pipelines.BinaryClassificationPipeline* method), 1129

compute_estimator_features() (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1072

compute_estimator_features() (*evalml.pipelines.ClassificationPipeline* method), 1140

compute_estimator_features() (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1080

compute_estimator_features() (*evalml.pipelines.MulticlassClassificationPipeline* method), 1186

compute_estimator_features() (*evalml.pipelines.pipeline_base.PipelineBase* method), 1086

compute_estimator_features() (*evalml.pipelines.PipelineBase* method), 1196

compute_estimator_features() (*evalml.pipelines.regression_pipeline.RegressionPipeline* method), 1091

compute_estimator_features() (*evalml.pipelines.RegressionPipeline* method), 1207

compute_estimator_features() (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1096

compute_estimator_features() (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1102

compute_estimator_features() (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1108

compute_estimator_features() (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase* method), 1114

compute_estimator_features() (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* method), 1119

compute_estimator_features() (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* method), 1234

compute_estimator_features() (*evalml.pipelines.TimeSeriesClassificationPipeline* method), 1240

compute_estimator_features() (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1245

compute_estimator_features() (*evalml.pipelines.TimeSeriesRegressionPipeline* method), 1251

compute_final_component_features() (*evalml.pipelines.component_graph.ComponentGraph* method), 1076

compute_final_component_features() (*evalml.pipelines.ComponentGraph* method), 1144

compute_order() (*evalml.pipelines.component_graph.ComponentGraph* property), 1077

compute_order() (*evalml.pipelines.ComponentGraph* property), 1144

ConfusionMatrix (class in *evalml.model_understanding*), 293

confusion_matrix() (in module *evalml.model_understanding.graphs*), 281

convert_to_seconds() (in module *evalml.utils*), 1296

convert_to_seconds() (in module *evalml.utils.gen_utils*), 1291

CostBenefitMatrix (class in *evalml.objectives*), 410

CostBenefitMatrix (class in *evalml.objectives.cost_benefit_matrix*), 306

create_objectives() (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* static method), 1066

create_objectives() (*evalml.pipelines.BinaryClassificationPipeline* static method), 1129

create_objectives() (*evalml.pipelines.classification_pipeline.ClassificationPipeline* static method), 1072

create_objectives() (*evalml.pipelines.ClassificationPipeline* static method), 1140

create_objectives() (*evalml.pipelines.MulticlassClassificationPipeline* static method), 1144

[static method](#)), 1081
[create_objectives\(\)](#) ([evalml.pipelines.MulticlassClassificationPipeline](#) [static method](#)), 1187
[create_objectives\(\)](#) ([evalml.pipelines.pipeline_base.PipelineBase](#) [static method](#)), 1086
[create_objectives\(\)](#) ([evalml.pipelines.PipelineBase](#) [static method](#)), 1196
[create_objectives\(\)](#) ([evalml.pipelines.regression_pipeline.RegressionPipeline](#) [static method](#)), 1091
[create_objectives\(\)](#) ([evalml.pipelines.RegressionPipeline](#) [static method](#)), 1208
[create_objectives\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) [static method](#)), 1096
[create_objectives\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline](#) [static method](#)), 1103
[create_objectives\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline](#) [static method](#)), 1108
[create_objectives\(\)](#) ([evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase](#) [static method](#)), 1114
[create_objectives\(\)](#) ([evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline](#) [static method](#)), 1119
[create_objectives\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) [static method](#)), 1234
[create_objectives\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#) [static method](#)), 1240
[create_objectives\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#) [static method](#)), 1246
[create_objectives\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#) [static method](#)), 1251
[cross_entropy\(\)](#) (in [module](#) [evalml.model_understanding.prediction_explanations.explainers](#)), 272
[custom_name\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) [property](#)), 1066
[custom_name\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) [property](#)), 1129
[custom_name\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) [property](#)), 1072
[custom_name\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) [property](#)), 1140
[custom_name\(\)](#) ([evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline](#) [property](#)), 1081
[custom_name\(\)](#) ([evalml.pipelines.MulticlassClassificationPipeline](#) [property](#)), 1187
[custom_name\(\)](#) ([evalml.pipelines.pipeline_base.PipelineBase](#) [property](#)), 1086
[custom_name\(\)](#) ([evalml.pipelines.PipelineBase](#) [property](#)), 1196
[custom_name\(\)](#) ([evalml.pipelines.regression_pipeline.RegressionPipeline](#) [property](#)), 1091
[custom_name\(\)](#) ([evalml.pipelines.RegressionPipeline](#) [property](#)), 1208
[custom_name\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) [property](#)), 1096
[custom_name\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline](#) [property](#)), 1103
[custom_name\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline](#) [property](#)), 1114
[custom_name\(\)](#) ([evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase](#) [property](#)), 1119
[custom_name\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) [property](#)), 1234
[custom_name\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#) [property](#)), 1240
[custom_name\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#) [property](#)), 1246
[custom_name\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#) [property](#)), 1251

D

[DataComputation](#) (class in [evalml.automl.engine.dask_engine](#)), 179
[DaskEngine](#) (class in [evalml.automl.engine](#)), 187
[DaskEngine](#) (class in [evalml.automl.engine.dask_engine](#)), 180
[DataCheck](#) (class in [evalml.data_checks](#)), 242
[DataCheck](#) (class in [evalml.data_checks.data_check](#)), 214
[DataCheckAction](#) (class in [evalml.data_checks](#)), 243
[DataCheckAction](#) (class in [evalml.data_checks.data_check_action](#)), 215
[DataCheckActionCode](#) (class in [evalml.data_checks](#)), 243
[DataCheckActionCode](#) (class in [evalml.data_checks.data_check_action_code](#)), 215
[DataCheckError](#) (class in [evalml.data_checks](#)), 243
[DataCheckError](#) (class in [evalml.data_checks.data_check_message](#)), 216
[DataCheckInitError](#), 265, 267

DataCheckMessage (class in <i>evalml.data_checks</i>), 244	decision_function() (<i>evalml.objectives.binary_classification_objective.BinaryClassificationObjective</i> method), 304
DataCheckMessage (class in <i>evalml.data_checks.data_check_message</i>), 216	decision_function() (<i>evalml.objectives.BinaryClassificationObjective</i> method), 408
DataCheckMessageCode (class in <i>evalml.data_checks</i>), 244	decision_function() (<i>evalml.objectives.cost_benefit_matrix.CostBenefitMatrix</i> method), 307
DataCheckMessageCode (class in <i>evalml.data_checks.data_check_message_code</i>), 217	decision_function() (<i>evalml.objectives.CostBenefitMatrix</i> method), 411
DataCheckMessageType (class in <i>evalml.data_checks</i>), 245	decision_function() (<i>evalml.objectives.F1</i> method), 415
DataCheckMessageType (class in <i>evalml.data_checks.data_check_message_type</i>), 219	decision_function() (<i>evalml.objectives.fraud_cost.FraudCost</i> method), 310
DataChecks (class in <i>evalml.data_checks</i>), 246	decision_function() (<i>evalml.objectives.FraudCost</i> method), 422
DataChecks (class in <i>evalml.data_checks.data_checks</i>), 220	decision_function() (<i>evalml.objectives.Gini</i> method), 425
DataCheckWarning (class in <i>evalml.data_checks</i>), 246	decision_function() (<i>evalml.objectives.lead_scoring.LeadScoring</i> method), 313
DataCheckWarning (class in <i>evalml.data_checks.data_check_message</i>), 217	decision_function() (<i>evalml.objectives.LeadScoring</i> method), 427
DateTimeFeaturizer (class in <i>evalml.pipelines.components</i>), 942	decision_function() (<i>evalml.objectives.LogLossBinary</i> method), 430
DateTimeFeaturizer (class in <i>evalml.pipelines.components.transformers</i>), 851	decision_function() (<i>evalml.objectives.MCCBinary</i> method), 428
DateTimeFeaturizer (class in <i>evalml.pipelines.components.transformers.preprocessing</i>), 791	decision_function() (<i>evalml.objectives.Precision</i> method), 451
DateTimeFeaturizer (class in <i>evalml.pipelines.components.transformers.preprocessing.datetime_featurizer</i>), 760	decision_function() (<i>evalml.objectives.Recall</i> method), 460
DateTimeFormatDataCheck (class in <i>evalml.data_checks</i>), 246	decision_function() (<i>evalml.objectives.sensitivity_low_alert.SensitivityLowAlert</i> method), 323
DateTimeFormatDataCheck (class in <i>evalml.data_checks.datetime_format_data_check</i>), 221	decision_function() (<i>evalml.objectives.SensitivityLowAlert</i> method), 473
DateTimeNaNDataCheck (class in <i>evalml.data_checks</i>), 247	decision_function() (<i>evalml.objectives.standard_metrics.AccuracyBinary</i> method), 326
DateTimeNaNDataCheck (class in <i>evalml.data_checks.datetime_nan_data_check</i>), 222	decision_function() (<i>evalml.objectives.standard_metrics.AUC</i> method), 330
debug() (<i>evalml.automl.engine.engine_base.JobLogger</i> method), 183	decision_function() (<i>evalml.objectives.standard_metrics.BalancedAccuracyBinary</i> method), 337
decision_function() (<i>evalml.objectives.AccuracyBinary</i> method), 393	decision_function() (<i>evalml.objectives.standard_metrics.F1</i> method), 404
decision_function() (<i>evalml.objectives.AUC</i> method), 397	
decision_function() (<i>evalml.objectives.BalancedAccuracyBinary</i> method), 404	

method), 342

`decision_function()`
(*evalml.objectives.standard_metrics.Gini*
method), 349

`decision_function()`
(*evalml.objectives.standard_metrics.LogLossBinary*
method), 351

`decision_function()`
(*evalml.objectives.standard_metrics.MCCBinary*
method), 360

`decision_function()`
(*evalml.objectives.standard_metrics.Precision*
method), 369

`decision_function()`
(*evalml.objectives.standard_metrics.Recall*
method), 377

`decision_tree_data_from_estimator()` (in
module *evalml.model_understanding.graphs*),
281

`decision_tree_data_from_pipeline()` (in
module *evalml.model_understanding.graphs*),
281

`DecisionTreeClassifier` (class in
evalml.pipelines), 1146

`DecisionTreeClassifier` (class in
evalml.pipelines.components), 944

`DecisionTreeClassifier` (class in
evalml.pipelines.components.estimators),
644

`DecisionTreeClassifier` (class in
evalml.pipelines.components.estimators.classifiers),
529

`DecisionTreeClassifier` (class in
evalml.pipelines.components.estimators.classifiers.decision_tree_classifier),
497

`DecisionTreeRegressor` (class in
evalml.pipelines), 1148

`DecisionTreeRegressor` (class in
evalml.pipelines.components), 947

`DecisionTreeRegressor` (class in
evalml.pipelines.components.estimators),
647

`DecisionTreeRegressor` (class in
evalml.pipelines.components.estimators.regressors),
602

`DecisionTreeRegressor` (class in
evalml.pipelines.components.estimators.regressors.decision_tree_regressor),
565

`DEFAULT_METRICS` (in module
evalml.model_understanding.prediction_explanations.explainers),
272

`default_parameters()`
(*evalml.pipelines.ARIMARegressor* *method*),
1126

`default_parameters()`
(*evalml.pipelines.CatBoostClassifier* *method*),
1134

`default_parameters()`
(*evalml.pipelines.CatBoostRegressor* *method*),
1137

`default_parameters()`
(*evalml.pipelines.component_graph.ComponentGraph*
property), 1077

`default_parameters()`
(*evalml.pipelines.ComponentGraph* *prop-*
erty), 1144

`default_parameters()`
(*evalml.pipelines.components.ARIMARegressor*
method), 928

`default_parameters()`
(*evalml.pipelines.components.BaselineClassifier*
method), 930

`default_parameters()`
(*evalml.pipelines.components.BaselineRegressor*
method), 933

`default_parameters()`
(*evalml.pipelines.components.CatBoostClassifier*
method), 935

`default_parameters()`
(*evalml.pipelines.components.CatBoostRegressor*
method), 938

`default_parameters()`
(*evalml.pipelines.components.component_base.ComponentBase*
method), 918

`default_parameters()`
(*evalml.pipelines.components.ComponentBase*
method), 940

`default_parameters()`
(*evalml.pipelines.components.DateFeatureTransformer*
method), 943

`default_parameters()`
(*evalml.pipelines.components.DecisionTreeClassifier*
method), 945

`default_parameters()`
(*evalml.pipelines.components.DecisionTreeRegressor*
method), 948

`default_parameters()`
(*evalml.pipelines.components.DelayedFeatureTransformer*
method), 951

`default_parameters()`
(*evalml.pipelines.components.DFSTransformer*
method), 953

`default_parameters()`
(*evalml.pipelines.components.DropColumns*
method), 955

`default_parameters()`
(*evalml.pipelines.components.DropNullColumns*
method), 957

<code>default_parameters()</code> (<i>evalml.pipelines.components.DropRowsTransformer</i> <i>method</i>), 959	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.catboost_classifier</i> <i>method</i>), 495
<code>default_parameters()</code> (<i>evalml.pipelines.components.ElasticNetClassifier</i> <i>method</i>), 962	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.CatBoostClassifier</i> <i>method</i>), 528
<code>default_parameters()</code> (<i>evalml.pipelines.components.ElasticNetRegressor</i> <i>method</i>), 964	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.decision_tree_classifier</i> <i>method</i>), 498
<code>default_parameters()</code> (<i>evalml.pipelines.components.EmailFeaturizer</i> <i>method</i>), 966	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier</i> <i>method</i>), 531
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier</i> <i>method</i>), 476	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.elasticnet_classifier</i> <i>method</i>), 501
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier</i> <i>method</i>), 479	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier</i> <i>method</i>), 534
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_regressor</i> <i>method</i>), 481	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier</i> <i>method</i>), 505
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier</i> <i>method</i>), 484	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier</i> <i>method</i>), 537
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier</i> <i>method</i>), 487	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.kneighbors_classifier</i> <i>method</i>), 508
<code>default_parameters()</code> (<i>evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor</i> <i>method</i>), 489	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier</i> <i>method</i>), 540
<code>default_parameters()</code> (<i>evalml.pipelines.components.Estimator</i> <i>method</i>), 968	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.lightgbm_classifier</i> <i>method</i>), 511
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.ARIMAREgressor</i> <i>method</i>), 633	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.LightGBMClassifier</i> <i>method</i>), 543
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.BaselineClassifier</i> <i>method</i>), 635	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier</i> <i>method</i>), 514
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.BaselineRegressor</i> <i>method</i>), 638	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier</i> <i>method</i>), 546
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.CatBoostClassifier</i> <i>method</i>), 640	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.RandomForestClassifier</i> <i>method</i>), 549
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.CatBoostRegressor</i> <i>method</i>), 643	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.rf_classifier</i> <i>method</i>), 517
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.baseline_classifier</i> <i>method</i>), 492	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.svm_classifier</i> <i>method</i>), 520
<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.BaselineClassifier</i> <i>method</i>), 525	<code>default_parameters()</code> (<i>evalml.pipelines.components.estimators.classifiers.SVMClassifier</i> <i>method</i>), 551

default_parameters() (evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier, method), 522	default_parameters() (evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier, method), 557
default_parameters() (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier, method), 554	default_parameters() (evalml.pipelines.components.estimators.regressors.ARIMARegressor, method), 596
default_parameters() (evalml.pipelines.components.estimators.DecisionTreeClassifier, method), 646	default_parameters() (evalml.pipelines.components.estimators.regressors.baseline_regressor.BaselineRegressor, method), 560
default_parameters() (evalml.pipelines.components.estimators.DecisionTreeRegressor, method), 649	default_parameters() (evalml.pipelines.components.estimators.regressors.BaselineRegressor, method), 598
default_parameters() (evalml.pipelines.components.estimators.ElasticNetClassifier, method), 652	default_parameters() (evalml.pipelines.components.estimators.regressors.catboost_regressor.CatBoostRegressor, method), 563
default_parameters() (evalml.pipelines.components.estimators.ElasticNetRegressor, method), 655	default_parameters() (evalml.pipelines.components.estimators.regressors.CatBoostRegressor, method), 600
default_parameters() (evalml.pipelines.components.estimators.Estimator, method), 657	default_parameters() (evalml.pipelines.components.estimators.regressors.decision_tree.DecisionTreeRegressor, method), 566
default_parameters() (evalml.pipelines.components.estimators.estimator.Estimator, method), 629	default_parameters() (evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor, method), 603
default_parameters() (evalml.pipelines.components.estimators.ExtraTreesClassifier, method), 660	default_parameters() (evalml.pipelines.components.estimators.regressors.elasticnet_regressor.ElasticNetRegressor, method), 569
default_parameters() (evalml.pipelines.components.estimators.ExtraTreesRegressor, method), 663	default_parameters() (evalml.pipelines.components.estimators.regressors.ElasticNetRegressor, method), 606
default_parameters() (evalml.pipelines.components.estimators.KNeighborsClassifier, method), 666	default_parameters() (evalml.pipelines.components.estimators.regressors.et_regressor.ExtraTreesRegressor, method), 572
default_parameters() (evalml.pipelines.components.estimators.LightGBMClassifier, method), 668	default_parameters() (evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor, method), 609
default_parameters() (evalml.pipelines.components.estimators.LightGBMRegressor, method), 671	default_parameters() (evalml.pipelines.components.estimators.regressors.lightgbm_regressor.LightGBMRegressor, method), 575
default_parameters() (evalml.pipelines.components.estimators.LinearRegressor, method), 674	default_parameters() (evalml.pipelines.components.estimators.regressors.LightGBMRegressor, method), 612
default_parameters() (evalml.pipelines.components.estimators.LogisticRegressionClassifier, method), 676	default_parameters() (evalml.pipelines.components.estimators.regressors.linear_regressor.LinearRegressor, method), 578
default_parameters() (evalml.pipelines.components.estimators.ProphetRegressor, method), 679	default_parameters() (evalml.pipelines.components.estimators.regressors.LinearRegressor, method), 614
default_parameters() (evalml.pipelines.components.estimators.RandomForestClassifier, method), 681	default_parameters() (evalml.pipelines.components.estimators.regressors.prophet_regressor.ProphetRegressor, method), 581
default_parameters() (evalml.pipelines.components.estimators.RandomForestRegressor, method), 683	default_parameters() (evalml.pipelines.components.estimators.regressors.ProphetRegressor, method), 617

<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.RandomForestRegressor</code> <code>method</code>), 619	<code>default_parameters()</code> (<code>evalml.pipelines.components.LightGBMClassifier</code> <code>method</code>), 984
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.rf_regressor</code> <code>method</code>), 584	<code>default_parameters()</code> (<code>evalml.pipelines.components.LightGBMRegressor</code> <code>method</code>), 987
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.svm_regressor</code> <code>method</code>), 587	<code>default_parameters()</code> (<code>evalml.pipelines.components.LinearDiscriminantAnalysis</code> <code>method</code>), 989
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.SVMRegressor</code> <code>method</code>), 622	<code>default_parameters()</code> (<code>evalml.pipelines.components.LinearRegressor</code> <code>method</code>), 992
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.time_series_baseline_estimator</code> <code>method</code>), 589	<code>default_parameters()</code> (<code>evalml.pipelines.components.LogisticRegressionClassifier</code> <code>method</code>), 994
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator</code> <code>method</code>), 624	<code>default_parameters()</code> (<code>evalml.pipelines.components.LogTransformer</code> <code>method</code>), 996
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.xgboost_regressor</code> <code>method</code>), 592	<code>default_parameters()</code> (<code>evalml.pipelines.components.LSA</code> <code>method</code>), 999
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.XGBoostRegressor</code> <code>method</code>), 626	<code>default_parameters()</code> (<code>evalml.pipelines.components.OneHotEncoder</code> <code>method</code>), 1001
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.SVMClassifier</code> <code>method</code>), 686	<code>default_parameters()</code> (<code>evalml.pipelines.components.Oversampler</code> <code>method</code>), 1004
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.SVMRegressor</code> <code>method</code>), 688	<code>default_parameters()</code> (<code>evalml.pipelines.components.PCA</code> <code>method</code>), 1006
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator</code> <code>method</code>), 690	<code>default_parameters()</code> (<code>evalml.pipelines.components.PerColumnImputer</code> <code>method</code>), 1008
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.XGBoostClassifier</code> <code>method</code>), 693	<code>default_parameters()</code> (<code>evalml.pipelines.components.PolynomialDetrender</code> <code>method</code>), 1011
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.XGBoostRegressor</code> <code>method</code>), 696	<code>default_parameters()</code> (<code>evalml.pipelines.components.ProphetRegressor</code> <code>method</code>), 1013
<code>default_parameters()</code> (<code>evalml.pipelines.components.ExtraTreesClassifier</code> <code>method</code>), 971	<code>default_parameters()</code> (<code>evalml.pipelines.components.RandomForestClassifier</code> <code>method</code>), 1016
<code>default_parameters()</code> (<code>evalml.pipelines.components.ExtraTreesRegressor</code> <code>method</code>), 974	<code>default_parameters()</code> (<code>evalml.pipelines.components.RandomForestRegressor</code> <code>method</code>), 1018
<code>default_parameters()</code> (<code>evalml.pipelines.components.FeatureSelector</code> <code>method</code>), 976	<code>default_parameters()</code> (<code>evalml.pipelines.components.RFClassifierSelectFromModel</code> <code>method</code>), 1020
<code>default_parameters()</code> (<code>evalml.pipelines.components.Imputer</code> <code>method</code>), 979	<code>default_parameters()</code> (<code>evalml.pipelines.components.RFRegressorSelectFromModel</code> <code>method</code>), 1023
<code>default_parameters()</code> (<code>evalml.pipelines.components.KNeighborsClassifier</code> <code>method</code>), 982	<code>default_parameters()</code> (<code>evalml.pipelines.components.SelectByType</code> <code>method</code>), 1025

`default_parameters()`
(*evalml.pipelines.components.SelectColumns*
method), 1027

`default_parameters()`
(*evalml.pipelines.components.SimpleImputer*
method), 1029

`default_parameters()`
(*evalml.pipelines.components.SklearnStackedEnsembleClassifier*
method), 1032

`default_parameters()`
(*evalml.pipelines.components.SklearnStackedEnsembleRegressor*
method), 1035

`default_parameters()`
(*evalml.pipelines.components.StandardScaler*
method), 1037

`default_parameters()`
(*evalml.pipelines.components.SVMClassifier*
method), 1039

`default_parameters()`
(*evalml.pipelines.components.SVMRegressor*
method), 1041

`default_parameters()`
(*evalml.pipelines.components.TargetEncoder*
method), 1044

`default_parameters()`
(*evalml.pipelines.components.TargetImputer*
method), 1046

`default_parameters()`
(*evalml.pipelines.components.TextFeaturizer*
method), 1048

`default_parameters()`
(*evalml.pipelines.components.TimeSeriesBaselineEstimator*
method), 1051

`default_parameters()`
(*evalml.pipelines.components.Transformer*
method), 1053

`default_parameters()`
(*evalml.pipelines.components.transformers.column_selector.ColumnSelector*
method), 837

`default_parameters()`
(*evalml.pipelines.components.transformers.column_selector.DropColumns*
method), 839

`default_parameters()`
(*evalml.pipelines.components.transformers.column_selector.SelectByIndex*
method), 841

`default_parameters()`
(*evalml.pipelines.components.transformers.column_selector.SelectByType*
method), 843

`default_parameters()`
(*evalml.pipelines.components.transformers.DateTimeFeaturizer*
method), 852

`default_parameters()`
(*evalml.pipelines.components.transformers.DelayedFeatureTransformer*
method), 854

`default_parameters()`
(*evalml.pipelines.components.transformers.DFSTransformer*
method), 857

`default_parameters()`
(*evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimensionality*
method), 698

`default_parameters()`
(*evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimensionality*
method), 703

`default_parameters()`
(*evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimensionality*
method), 706

`default_parameters()`
(*evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimensionality*
method), 701

`default_parameters()`
(*evalml.pipelines.components.transformers.DropColumns*
method), 859

`default_parameters()`
(*evalml.pipelines.components.transformers.DropNullColumns*
method), 861

`default_parameters()`
(*evalml.pipelines.components.transformers.DropRowsTransformer*
method), 863

`default_parameters()`
(*evalml.pipelines.components.transformers.EmailFeaturizer*
method), 865

`default_parameters()`
(*evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder*
method), 709

`default_parameters()`
(*evalml.pipelines.components.transformers.encoders.OneHotEncoder*
method), 716

`default_parameters()`
(*evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder*
method), 713

`default_parameters()`
(*evalml.pipelines.components.transformers.encoders.TargetEncoder*
method), 719

`default_parameters()`
(*evalml.pipelines.components.transformers.feature_selection.FeatureSelector*
method), 721

`default_parameters()`
(*evalml.pipelines.components.transformers.feature_selection.FeatureSelector*
method), 731

`default_parameters()`
(*evalml.pipelines.components.transformers.feature_selection.rf_classifier.RandomForestClassifier*
method), 725

`default_parameters()`
(*evalml.pipelines.components.transformers.feature_selection.rf_classifier.RandomForestClassifier*
method), 728

`default_parameters()`
(*evalml.pipelines.components.transformers.feature_selection.RFClassifier*
method), 734

<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.feature_selection.RFRegressorSelectFromModel</code> method), 736	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.PolynomialDetrender</code> method), 888
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.FeatureSelector</code> method), 867	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.datetime</code> method), 761
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.Imputer</code> method), 870	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DateTim</code> method), 792
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.Imputer</code> method), 751	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.delayed</code> method), 764
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.imputer</code> method), 739	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.Delayed</code> method), 794
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.per_column</code> method), 742	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DFSTra</code> method), 797
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.PerColumnImputer</code> method), 753	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.drop_n</code> method), 767
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.simple</code> method), 745	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.drop_re</code> method), 769
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.SimpleImputer</code> method), 756	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DropN</code> method), 799
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.target</code> method), 748	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DropRo</code> method), 801
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.imputers.TargetImputer</code> method), 758	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.EmailF</code> method), 803
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.LinearDiscriminantAnalysis</code> method), 872	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.feature</code> method), 772
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.LogTransformer</code> method), 874	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.log_tra</code> method), 774
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.LSA</code> method), 876	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.LogTra</code> method), 805
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.OneHotEncoder</code> method), 879	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.LSA</code> method), 807
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.Oversampler</code> method), 882	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.lsa.LSA</code> method), 777
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.PCA</code> method), 884	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.polynom</code> method), 779
<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.PerColumnImputer</code> method), 886	<code>default_parameters()</code> (<code>evalml.pipelines.components.transformers.preprocessing.Polynom</code> method), 809

default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 782	default_parameters() (evalml.pipelines.components.transformers.SimpleImputer method), 900
default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 784	default_parameters() (evalml.pipelines.components.transformers.StandardScaler method), 902
default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 812	default_parameters() (evalml.pipelines.components.transformers.TargetEncoder method), 905
default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 814	default_parameters() (evalml.pipelines.components.transformers.TargetImputer method), 907
default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 787	default_parameters() (evalml.pipelines.components.transformers.FeatureTransformer method), 909
default_parameters() (evalml.pipelines.components.transformers.preprocessing.text.MultiTextFeaturizer method), 789	default_parameters() (evalml.pipelines.components.transformers.URLFeaturizer method), 912
default_parameters() (evalml.pipelines.components.transformers.preprocessing.URLFeaturizer method), 816	default_parameters() (evalml.pipelines.components.transformer.TargetTransformer method), 846
default_parameters() (evalml.pipelines.components.transformers.RFClassifierSelector method), 891	default_parameters() (evalml.pipelines.components.transformer.Transformers method), 848
default_parameters() (evalml.pipelines.components.transformers.RFRegressorSelector method), 894	default_parameters() (evalml.pipelines.components.transformers.Undersampler method), 914
default_parameters() (evalml.pipelines.components.transformers.samplers.base_sampler method), 819	default_parameters() (evalml.pipelines.components.transformers.URLFeaturizer method), 916
default_parameters() (evalml.pipelines.components.transformers.samplers.Oversampler method), 827	default_parameters() (evalml.pipelines.components.Undersampler method), 1055
default_parameters() (evalml.pipelines.components.transformers.samplers.oversampler method), 822	default_parameters() (evalml.pipelines.components.URLFeaturizer method), 1057
default_parameters() (evalml.pipelines.components.transformers.samplers.Undersampler method), 830	default_parameters() (evalml.pipelines.components.XGBoostClassifier method), 1060
default_parameters() (evalml.pipelines.components.transformers.samplers.undersampler method), 824	default_parameters() (evalml.pipelines.components.XGBoostRegressor method), 1062
default_parameters() (evalml.pipelines.components.transformers.scalers.standard_scaler method), 832	default_parameters() (evalml.pipelines.components.DecisionTreeClassifier method), 1147
default_parameters() (evalml.pipelines.components.transformers.scalers.StandardScaler method), 834	default_parameters() (evalml.pipelines.DecisionTreeRegressor method), 1150
default_parameters() (evalml.pipelines.components.transformers.SelectByType method), 896	default_parameters() (evalml.pipelines.DelayedFeatureTransformer method), 1153
default_parameters() (evalml.pipelines.components.transformers.SelectColumns method), 898	default_parameters() (evalml.pipelines.DFSTransformer method), 1155

[default_parameters\(\)](#) ([evalml.pipelines.ElasticNetClassifier](#) method), 1158
[default_parameters\(\)](#) ([evalml.pipelines.ElasticNetRegressor](#) method), 1160
[default_parameters\(\)](#) ([evalml.pipelines.Estimator](#) method), 1162
[default_parameters\(\)](#) ([evalml.pipelines.ExtraTreesClassifier](#) method), 1165
[default_parameters\(\)](#) ([evalml.pipelines.ExtraTreesRegressor](#) method), 1168
[default_parameters\(\)](#) ([evalml.pipelines.FeatureSelector](#) method), 1170
[default_parameters\(\)](#) ([evalml.pipelines.KNeighborsClassifier](#) method), 1173
[default_parameters\(\)](#) ([evalml.pipelines.LightGBMClassifier](#) method), 1176
[default_parameters\(\)](#) ([evalml.pipelines.LightGBMRegressor](#) method), 1179
[default_parameters\(\)](#) ([evalml.pipelines.LinearRegressor](#) method), 1181
[default_parameters\(\)](#) ([evalml.pipelines.LogisticRegressionClassifier](#) method), 1183
[default_parameters\(\)](#) ([evalml.pipelines.OneHotEncoder](#) method), 1191
[default_parameters\(\)](#) ([evalml.pipelines.PerColumnImputer](#) method), 1193
[default_parameters\(\)](#) ([evalml.pipelines.ProphetRegressor](#) method), 1200
[default_parameters\(\)](#) ([evalml.pipelines.RandomForestClassifier](#) method), 1202
[default_parameters\(\)](#) ([evalml.pipelines.RandomForestRegressor](#) method), 1205
[default_parameters\(\)](#) ([evalml.pipelines.RFClassifierSelectFromModel](#) method), 1212
[default_parameters\(\)](#) ([evalml.pipelines.RFRegressorSelectFromModel](#) method), 1214
[default_parameters\(\)](#) ([evalml.pipelines.SimpleImputer](#) method), 1216
[default_parameters\(\)](#) ([evalml.pipelines.SklearnStackedEnsembleClassifier](#) method), 1219
[default_parameters\(\)](#) ([evalml.pipelines.SklearnStackedEnsembleRegressor](#) method), 1222
[default_parameters\(\)](#) ([evalml.pipelines.StandardScaler](#) method), 1224
[default_parameters\(\)](#) ([evalml.pipelines.SVMClassifier](#) method), 1226
[default_parameters\(\)](#) ([evalml.pipelines.SVMRegressor](#) method), 1228
[default_parameters\(\)](#) ([evalml.pipelines.TargetEncoder](#) method), 1231
[default_parameters\(\)](#) ([evalml.pipelines.Transformer](#) method), 1255
[default_parameters\(\)](#) ([evalml.pipelines.XGBoostClassifier](#) method), 1257
[default_parameters\(\)](#) ([evalml.pipelines.XGBoostRegressor](#) method), 1260
[DefaultAlgorithm](#) (class in [evalml.automl.automl_algorithm](#)), 173
[DefaultAlgorithm](#) (class in [evalml.automl.automl_algorithm.default_algorithm](#)), 168
[DefaultDataChecks](#) (class in [evalml.data_checks](#)), 248
[DefaultDataChecks](#) (class in [evalml.data_checks.default_data_checks](#)), 223
[DelayedFeatureTransformer](#) (class in [evalml.pipelines](#)), 1151
[DelayedFeatureTransformer](#) (class in [evalml.pipelines.components](#)), 950
[DelayedFeatureTransformer](#) (class in [evalml.pipelines.components.transformers](#)), 853
[DelayedFeatureTransformer](#) (class in [evalml.pipelines.components.transformers.preprocessing](#)), 793
[DelayedFeatureTransformer](#) (class in [evalml.pipelines.components.transformers.preprocessing.delayed](#)), 763
[deprecate_arg\(\)](#) (in module [evalml.utils](#)), 1296
[deprecate_arg\(\)](#) (in module [evalml.utils.gen_utils](#)), 1291

`describe()` (`evalml.pipelines.ARIMAREgressor` `describe()` (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble`
`method`), 1126 `method`), 476
`describe()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline`
`method`), 1066 `method`), 479
`describe()` (`evalml.pipelines.BinaryClassificationPipeline` `describe()` (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble`
`method`), 1130 `method`), 481
`describe()` (`evalml.pipelines.CatBoostClassifier` `describe()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsemble`
`method`), 1134 `method`), 484
`describe()` (`evalml.pipelines.CatBoostRegressor` `describe()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsemble`
`method`), 1137 `method`), 487
`describe()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` `describe()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsemble`
`method`), 1072 `method`), 490
`describe()` (`evalml.pipelines.ClassificationPipeline` `describe()` (`evalml.pipelines.components.Estimator`
`method`), 1140 `method`), 969
`describe()` (`evalml.pipelines.component_graph.ComponentGraph` `describe()` (`evalml.pipelines.components.estimators.ARIMAREgressor`
`method`), 1077 `method`), 633
`describe()` (`evalml.pipelines.ComponentGraph` `describe()` (`evalml.pipelines.components.estimators.BaselineClassifier`
`method`), 1144 `method`), 636
`describe()` (`evalml.pipelines.components.ARIMAREgressor` `describe()` (`evalml.pipelines.components.estimators.BaselineRegressor`
`method`), 928 `method`), 638
`describe()` (`evalml.pipelines.components.BaselineClassifier` `describe()` (`evalml.pipelines.components.estimators.CatBoostClassifier`
`method`), 931 `method`), 641
`describe()` (`evalml.pipelines.components.BaselineRegressor` `describe()` (`evalml.pipelines.components.estimators.CatBoostRegressor`
`method`), 933 `method`), 643
`describe()` (`evalml.pipelines.components.CatBoostClassifier` `describe()` (`evalml.pipelines.components.estimators.classifiers.baseline`
`method`), 935 `method`), 492
`describe()` (`evalml.pipelines.components.CatBoostRegressor` `describe()` (`evalml.pipelines.components.estimators.classifiers.Baseline`
`method`), 938 `method`), 525
`describe()` (`evalml.pipelines.components.component_base.ComponentBase` `describe()` (`evalml.pipelines.components.estimators.classifiers.catboost`
`method`), 919 `method`), 495
`describe()` (`evalml.pipelines.components.ComponentBase` `describe()` (`evalml.pipelines.components.estimators.classifiers.CatBoost`
`method`), 940 `method`), 528
`describe()` (`evalml.pipelines.components.DateTimeFeaturizer` `describe()` (`evalml.pipelines.components.estimators.classifiers.decision`
`method`), 943 `method`), 498
`describe()` (`evalml.pipelines.components.DecisionTreeClassifier` `describe()` (`evalml.pipelines.components.estimators.classifiers.Decision`
`method`), 946 `method`), 531
`describe()` (`evalml.pipelines.components.DecisionTreeRegressor` `describe()` (`evalml.pipelines.components.estimators.classifiers.elasticnet`
`method`), 948 `method`), 502
`describe()` (`evalml.pipelines.components.DelayedFeatureTransformer` `describe()` (`evalml.pipelines.components.estimators.classifiers.ElasticNet`
`method`), 951 `method`), 534
`describe()` (`evalml.pipelines.components.DFSTransformer` `describe()` (`evalml.pipelines.components.estimators.classifiers.et_classifier`
`method`), 953 `method`), 505
`describe()` (`evalml.pipelines.components.DropColumns` `describe()` (`evalml.pipelines.components.estimators.classifiers.ExtraTree`
`method`), 955 `method`), 537
`describe()` (`evalml.pipelines.components.DropNullColumns` `describe()` (`evalml.pipelines.components.estimators.classifiers.kneighbors`
`method`), 957 `method`), 508
`describe()` (`evalml.pipelines.components.DropRowsTransformer` `describe()` (`evalml.pipelines.components.estimators.classifiers.KNeighbors`
`method`), 959 `method`), 540
`describe()` (`evalml.pipelines.components.ElasticNetClassifier` `describe()` (`evalml.pipelines.components.estimators.classifiers.lightgbm`
`method`), 962 `method`), 511
`describe()` (`evalml.pipelines.components.ElasticNetRegressor` `describe()` (`evalml.pipelines.components.estimators.classifiers.LightGBM`
`method`), 964 `method`), 543
`describe()` (`evalml.pipelines.components.EmailFeaturizer` `describe()` (`evalml.pipelines.components.estimators.classifiers.logistic`
`method`), 966 `method`), 514

`describe()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` method), 546
`describe()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` method), 563
`describe()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` method), 549
`describe()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` method), 601
`describe()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` method), 517
`describe()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` method), 566
`describe()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` method), 520
`describe()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` method), 603
`describe()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` method), 551
`describe()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` method), 569
`describe()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` method), 523
`describe()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` method), 606
`describe()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` method), 554
`describe()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` method), 572
`describe()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 646
`describe()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 609
`describe()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 649
`describe()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 575
`describe()` (`evalml.pipelines.components.estimators.ElasticNetClassifier` method), 652
`describe()` (`evalml.pipelines.components.estimators.ElasticNetClassifier` method), 612
`describe()` (`evalml.pipelines.components.estimators.ElasticNetRegressor` method), 655
`describe()` (`evalml.pipelines.components.estimators.ElasticNetRegressor` method), 578
`describe()` (`evalml.pipelines.components.estimators.Estimator` method), 657
`describe()` (`evalml.pipelines.components.estimators.Estimator` method), 614
`describe()` (`evalml.pipelines.components.estimators.estimator.Estimator` method), 629
`describe()` (`evalml.pipelines.components.estimators.estimator.Estimator` method), 581
`describe()` (`evalml.pipelines.components.estimators.ExtraTreesClassifier` method), 660
`describe()` (`evalml.pipelines.components.estimators.ExtraTreesClassifier` method), 617
`describe()` (`evalml.pipelines.components.estimators.ExtraTreesRegressor` method), 663
`describe()` (`evalml.pipelines.components.estimators.ExtraTreesRegressor` method), 619
`describe()` (`evalml.pipelines.components.estimators.KNeighborsClassifier` method), 666
`describe()` (`evalml.pipelines.components.estimators.KNeighborsClassifier` method), 584
`describe()` (`evalml.pipelines.components.estimators.LightGBMClassifier` method), 669
`describe()` (`evalml.pipelines.components.estimators.LightGBMClassifier` method), 587
`describe()` (`evalml.pipelines.components.estimators.LightGBMRegressor` method), 671
`describe()` (`evalml.pipelines.components.estimators.LightGBMRegressor` method), 622
`describe()` (`evalml.pipelines.components.estimators.LinearRegressor` method), 674
`describe()` (`evalml.pipelines.components.estimators.LinearRegressor` method), 589
`describe()` (`evalml.pipelines.components.estimators.LogisticRegressionClassifier` method), 676
`describe()` (`evalml.pipelines.components.estimators.LogisticRegressionClassifier` method), 624
`describe()` (`evalml.pipelines.components.estimators.ProphetRegressor` method), 679
`describe()` (`evalml.pipelines.components.estimators.ProphetRegressor` method), 592
`describe()` (`evalml.pipelines.components.estimators.RandomForestClassifier` method), 681
`describe()` (`evalml.pipelines.components.estimators.RandomForestClassifier` method), 626
`describe()` (`evalml.pipelines.components.estimators.RandomForestRegressor` method), 684
`describe()` (`evalml.pipelines.components.estimators.RandomForestRegressor` method), 686
`describe()` (`evalml.pipelines.components.estimators.regressor.ArimaRegressor` method), 557
`describe()` (`evalml.pipelines.components.estimators.regressor.ArimaRegressor` method), 688
`describe()` (`evalml.pipelines.components.estimators.regressor.ArimaRegressor` method), 596
`describe()` (`evalml.pipelines.components.estimators.regressor.ArimaRegressor` method), 691
`describe()` (`evalml.pipelines.components.estimators.regressor.baseline.LinearBaselineRegressor` method), 560
`describe()` (`evalml.pipelines.components.estimators.regressor.baseline.LinearBaselineRegressor` method), 693
`describe()` (`evalml.pipelines.components.estimators.regressor.baseline.RandomBaselineRegressor` method), 598
`describe()` (`evalml.pipelines.components.estimators.regressor.baseline.RandomBaselineRegressor` method), 696

`describe()` (`evalml.pipelines.components.ExtraTreesClassifier` method), 971
`describe()` (`evalml.pipelines.components.ExtraTreesRegressor` method), 974
`describe()` (`evalml.pipelines.components.FeatureSelector` method), 976
`describe()` (`evalml.pipelines.components.Imputer` method), 979
`describe()` (`evalml.pipelines.components.KNeighborsClassifier` method), 982
`describe()` (`evalml.pipelines.components.LightGBMClassifier` method), 985
`describe()` (`evalml.pipelines.components.LightGBMRegressor` method), 987
`describe()` (`evalml.pipelines.components.LinearDiscriminantAnalysis` method), 989
`describe()` (`evalml.pipelines.components.LinearRegressor` method), 992
`describe()` (`evalml.pipelines.components.LogisticRegressionClassifier` method), 995
`describe()` (`evalml.pipelines.components.LogTransformer` method), 997
`describe()` (`evalml.pipelines.components.LSA` method), 999
`describe()` (`evalml.pipelines.components.OneHotEncoder` method), 1002
`describe()` (`evalml.pipelines.components.Oversampler` method), 1004
`describe()` (`evalml.pipelines.components.PCA` method), 1006
`describe()` (`evalml.pipelines.components.PerColumnImputer` method), 1009
`describe()` (`evalml.pipelines.components.PolynomialDegreeReducer` method), 1011
`describe()` (`evalml.pipelines.components.ProphetRegressor` method), 1013
`describe()` (`evalml.pipelines.components.RandomForestClassifier` method), 1016
`describe()` (`evalml.pipelines.components.RandomForestRegressor` method), 1018
`describe()` (`evalml.pipelines.components.RFClassifierSelector` method), 1020
`describe()` (`evalml.pipelines.components.RFRegressorSelector` method), 1023
`describe()` (`evalml.pipelines.components.SelectByType` method), 1025
`describe()` (`evalml.pipelines.components.SelectColumns` method), 1027
`describe()` (`evalml.pipelines.components.SimpleImputer` method), 1030
`describe()` (`evalml.pipelines.components.SklearnStackedEnsembleClassifier` method), 1032
`describe()` (`evalml.pipelines.components.SklearnStackedEnsembleRegressor` method), 1035
`describe()` (`evalml.pipelines.components.StandardScaler` method), 1037
`describe()` (`evalml.pipelines.components.SVMClassifier` method), 1039
`describe()` (`evalml.pipelines.components.SVMRegressor` method), 1042
`describe()` (`evalml.pipelines.components.TargetEncoder` method), 1044
`describe()` (`evalml.pipelines.components.TargetImputer` method), 1046
`describe()` (`evalml.pipelines.components.TextFeaturizer` method), 1048
`describe()` (`evalml.pipelines.components.TimeSeriesBaselineEstimator` method), 1051
`describe()` (`evalml.pipelines.components.Transformer` method), 1053
`describe()` (`evalml.pipelines.components.transformers.column_selector` method), 837
`describe()` (`evalml.pipelines.components.transformers.column_selector` method), 839
`describe()` (`evalml.pipelines.components.transformers.column_selector` method), 841
`describe()` (`evalml.pipelines.components.transformers.column_selector` method), 843
`describe()` (`evalml.pipelines.components.transformers.DateTimeFeaturizer` method), 852
`describe()` (`evalml.pipelines.components.transformers.DelayedFeature` method), 855
`describe()` (`evalml.pipelines.components.transformers.DFSTransformer` method), 857
`describe()` (`evalml.pipelines.components.transformers.dimensionality_` method), 698
`describe()` (`evalml.pipelines.components.transformers.dimensionality_` method), 703
`describe()` (`evalml.pipelines.components.transformers.dimensionality_` method), 706
`describe()` (`evalml.pipelines.components.transformers.dimensionality_` method), 701
`describe()` (`evalml.pipelines.components.transformers.DropColumns` method), 859
`describe()` (`evalml.pipelines.components.transformers.DropNullColumn` method), 861
`describe()` (`evalml.pipelines.components.transformers.DropRowsTrans` method), 863
`describe()` (`evalml.pipelines.components.transformers.EmailFeaturizer` method), 865
`describe()` (`evalml.pipelines.components.transformers.encoders.onehot` method), 709
`describe()` (`evalml.pipelines.components.transformers.encoders.OneHo` method), 716
`describe()` (`evalml.pipelines.components.transformers.encoders.target_` method), 713
`describe()` (`evalml.pipelines.components.transformers.encoders.Target` method), 719

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 722

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 731

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 725

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 728

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 734

`describe()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` transformers.preprocessing.D
method), 736

`describe()` (`evalml.pipelines.components.transformers.FeatureSelector` (evalml.pipelines.components.transformers.preprocessing.E
method), 868

`describe()` (`evalml.pipelines.components.transformers.InputTransformer` (evalml.pipelines.components.transformers.preprocessing.fe
method), 870

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.lo
method), 751

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.L
method), 739

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.L
method), 742

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.ls
method), 754

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.p
method), 745

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.P
method), 756

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.te
method), 748

`describe()` (`evalml.pipelines.components.transformers.input_transformer` (evalml.pipelines.components.transformers.preprocessing.te
method), 758

`describe()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis` (evalml.pipelines.components.transformers.preprocessing.Te
method), 872

`describe()` (`evalml.pipelines.components.transformers.LogTransformer` (evalml.pipelines.components.transformers.preprocessing.Te
method), 874

`describe()` (`evalml.pipelines.components.transformers.LSA` (evalml.pipelines.components.transformers.preprocessing.tr
method), 876

`describe()` (`evalml.pipelines.components.transformers.OneHotEncoder` (evalml.pipelines.components.transformers.preprocessing.tr
method), 879

`describe()` (`evalml.pipelines.components.transformers.Oversampler` (evalml.pipelines.components.transformers.preprocessing.U
method), 882

`describe()` (`evalml.pipelines.components.transformers.PCA` (evalml.pipelines.components.transformers.RFClassifierSele
method), 884

`describe()` (`evalml.pipelines.components.transformers.PreColumnInputTransformer` (evalml.pipelines.components.transformers.RFRegressorSele
method), 886

`describe()` (`evalml.pipelines.components.transformers.PolynomialDegree` (evalml.pipelines.components.transformers.samplers.base_s
method), 889

`describe()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` (evalml.pipelines.components.transformers.samplers.Overs
method), 761

`describe()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` (evalml.pipelines.components.transformers.samplers.oversa
method), 792

`describe()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` (evalml.pipelines.components.transformers.samplers.oversa
method), 764

`describe()` (`evalml.pipelines.components.transformers.samplers.BaseSampler` method), 825
`describe()` (`evalml.pipelines.components.transformers.samplers.BaseSampler` method), 1168
`describe()` (`evalml.pipelines.components.transformers.scalers.StandardScaler` method), 832
`describe()` (`evalml.pipelines.components.transformers.scalers.StandardScaler` method), 1170
`describe()` (`evalml.pipelines.components.transformers.scalers.StandardScaler` method), 835
`describe()` (`evalml.pipelines.components.transformers.scalers.StandardScaler` method), 1173
`describe()` (`evalml.pipelines.components.transformers.SelectByType` method), 897
`describe()` (`evalml.pipelines.components.transformers.SelectByType` method), 1176
`describe()` (`evalml.pipelines.components.transformers.SelectColumns` method), 898
`describe()` (`evalml.pipelines.components.transformers.SelectColumns` method), 1179
`describe()` (`evalml.pipelines.components.transformers.SimpleImputer` method), 901
`describe()` (`evalml.pipelines.components.transformers.SimpleImputer` method), 1181
`describe()` (`evalml.pipelines.components.transformers.StandardScaler` method), 903
`describe()` (`evalml.pipelines.components.transformers.StandardScaler` method), 1184
`describe()` (`evalml.pipelines.components.transformers.TargetEncoder` method), 905
`describe()` (`evalml.pipelines.components.transformers.TargetEncoder` method), 1081
`describe()` (`evalml.pipelines.components.transformers.TargetImputer` method), 907
`describe()` (`evalml.pipelines.components.transformers.TargetImputer` method), 1187
`describe()` (`evalml.pipelines.components.transformers.TargetFeaturizer` method), 910
`describe()` (`evalml.pipelines.components.transformers.TargetFeaturizer` method), 1191
`describe()` (`evalml.pipelines.components.transformers.Transformers` method), 912
`describe()` (`evalml.pipelines.components.transformers.Transformers` method), 1194
`describe()` (`evalml.pipelines.components.transformers.transformer_target_pipeline_base.PipelineBase` method), 846
`describe()` (`evalml.pipelines.components.transformers.transformer_target_pipeline_base.PipelineBase` method), 1086
`describe()` (`evalml.pipelines.components.transformers.transformer_target_pipeline_base.PipelineBase` method), 848
`describe()` (`evalml.pipelines.components.transformers.transformer_target_pipeline_base.PipelineBase` method), 1196
`describe()` (`evalml.pipelines.components.transformers.Undersampler` method), 914
`describe()` (`evalml.pipelines.components.transformers.Undersampler` method), 1200
`describe()` (`evalml.pipelines.components.transformers.URLFeaturizer` method), 916
`describe()` (`evalml.pipelines.components.transformers.URLFeaturizer` method), 1203
`describe()` (`evalml.pipelines.components.Undersampler` method), 1055
`describe()` (`evalml.pipelines.components.Undersampler` method), 1205
`describe()` (`evalml.pipelines.components.URLFeaturizer` method), 1057
`describe()` (`evalml.pipelines.components.URLFeaturizer` method), 1091
`describe()` (`evalml.pipelines.components.XGBoostClassifier` method), 1060
`describe()` (`evalml.pipelines.components.XGBoostClassifier` method), 1208
`describe()` (`evalml.pipelines.components.XGBoostRegressor` method), 1062
`describe()` (`evalml.pipelines.components.XGBoostRegressor` method), 1212
`describe()` (`evalml.pipelines.DecisionTreeClassifier` method), 1147
`describe()` (`evalml.pipelines.DecisionTreeClassifier` method), 1214
`describe()` (`evalml.pipelines.DecisionTreeRegressor` method), 1150
`describe()` (`evalml.pipelines.DecisionTreeRegressor` method), 1217
`describe()` (`evalml.pipelines.DelayedFeatureTransformer` method), 1153
`describe()` (`evalml.pipelines.DelayedFeatureTransformer` method), 1219
`describe()` (`evalml.pipelines.DFSTransformer` method), 1155
`describe()` (`evalml.pipelines.DFSTransformer` method), 1222
`describe()` (`evalml.pipelines.ElasticNetClassifier` method), 1158
`describe()` (`evalml.pipelines.ElasticNetClassifier` method), 1224
`describe()` (`evalml.pipelines.ElasticNetRegressor` method), 1160
`describe()` (`evalml.pipelines.ElasticNetRegressor` method), 1226
`describe()` (`evalml.pipelines.Estimator` method), 1162
`describe()` (`evalml.pipelines.Estimator` method), 1229
`describe()` (`evalml.pipelines.ExtraTreesClassifier` method), 1165
`describe()` (`evalml.pipelines.ExtraTreesClassifier` method), 1231

describe() (evalml.pipelines.time_series_classification_pipeline_base.TimeSeriesClassificationPipelineBase method), 1096

describe() (evalml.pipelines.time_series_classification_pipeline_base.TimeSeriesClassificationPipelineBase method), 1103

describe() (evalml.pipelines.time_series_classification_pipeline_base.TimeSeriesClassificationPipelineBase method), 1108

describe() (evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase method), 1114

describe() (evalml.pipelines.time_series_regression_pipeline_base.TimeSeriesRegressionPipelineBase method), 1119

describe() (evalml.pipelines.TimeSeriesBinaryClassificationPipeline method), 1234

describe() (evalml.pipelines.TimeSeriesClassificationPipeline method), 1240

describe() (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method), 1246

describe() (evalml.pipelines.TimeSeriesRegressionPipeline method), 1251

describe() (evalml.pipelines.Transformer method), 1255

describe() (evalml.pipelines.XGBoostClassifier method), 1257

describe() (evalml.pipelines.XGBoostRegressor method), 1260

describe_pipeline() (evalml.automl.automl_search.AutoMLSearch method), 195

describe_pipeline() (evalml.automl.AutoMLSearch method), 206

describe_pipeline() (evalml.AutoMLSearch method), 1302

detect_problem_type() (in module evalml.problem_types), 1276

detect_problem_type() (in module evalml.problem_types.utils), 1274

DFSTransformer (class in evalml.pipelines), 1154

DFSTransformer (class in evalml.pipelines.components), 952

DFSTransformer (class in evalml.pipelines.components.transformers), 856

DFSTransformer (class in evalml.pipelines.components.transformers.preprocessing), 796

DFSTransformer (class in evalml.pipelines.components.transformers.preprocessing.feature_tools), 771

done() (evalml.automl.engine.cf_engine.CFComputation method), 177

done() (evalml.automl.engine.dask_engine.DaskComputation method), 179

done() (evalml.automl.engine.engine_base.EngineComputation method), 182

DropColumns (class in evalml.pipelines.components.transformers.column_selectors), 838

DropNullColumns (class in evalml.pipelines.components), 956

DropNullColumns (class in evalml.pipelines.components.transformers), 860

DropNullColumns (class in evalml.pipelines.components.transformers.preprocessing), 798

DropNullColumns (class in evalml.pipelines.components.transformers.preprocessing.drop_rows_transformer), 766

DropRowsTransformer (class in evalml.pipelines.components), 958

DropRowsTransformer (class in evalml.pipelines.components.transformers), 862

DropRowsTransformer (class in evalml.pipelines.components.transformers.preprocessing), 800

DropRowsTransformer (class in evalml.pipelines.components.transformers.preprocessing.drop_rows_transformer), 768

E

ElasticNetClassifier (class in evalml.pipelines), 1156

ElasticNetClassifier (class in evalml.pipelines.components), 960

ElasticNetClassifier (class in evalml.pipelines.components.estimators), 480

ElasticNetClassifier (class in evalml.pipelines.components.estimators.classifiers), 532

ElasticNetClassifier (class in evalml.pipelines.components.estimators.classifiers.elasticnet_classifier), 500

ElasticNetRegressor (class in <i>evalml.pipelines</i>), 1159	<i>evalml.automl</i> module, 166
ElasticNetRegressor (class in <i>evalml.pipelines.components</i>), 963	<i>evalml.automl.automl_algorithm</i> module, 166
ElasticNetRegressor (class in <i>evalml.pipelines.components.estimators</i>), 653	<i>evalml.automl.automl_algorithm.automl_algorithm</i> module, 166
ElasticNetRegressor (class in <i>evalml.pipelines.components.estimators.regressors</i>), 605	<i>evalml.automl.automl_algorithm.default_algorithm</i> module, 167
ElasticNetRegressor (class in <i>evalml.pipelines.components.estimators.regressors.elasticnet_regressor</i>), 568	<i>evalml.automl.automl_algorithm.iterative_algorithm</i> module, 169
EmailFeaturizer (class in <i>evalml.pipelines.components</i>), 965	<i>evalml.automl.automl_search</i> module, 199
EmailFeaturizer (class in <i>evalml.pipelines.components.transformers</i>), 864	<i>evalml.automl.callbacks</i> module, 176
EmailFeaturizer (class in <i>evalml.pipelines.components.transformers.preprocessing</i>), 802	<i>evalml.automl.engine</i> module, 176
EmailFeaturizer (class in <i>evalml.pipelines.components.transformers.preprocessing.primitive_components</i>), 786	<i>evalml.automl.engine.cf_engine</i> module, 179
EmptyDataChecks (class in <i>evalml.data_checks</i>), 249	<i>evalml.automl.engine.dask_engine</i> module, 181
EmptyDataChecks (class in <i>evalml.data_checks.utils</i>), 239	<i>evalml.automl.engine.engine_base</i> module, 184
EngineBase (class in <i>evalml.automl</i>), 208	<i>evalml.automl.pipeline_search_plots</i> module, 200
EngineBase (class in <i>evalml.automl.engine</i>), 189	<i>evalml.automl.utils</i> module, 200
EngineBase (class in <i>evalml.automl.engine.engine_base</i>), 182	<i>evalml.data_checks</i> module, 212
EngineComputation (class in <i>evalml.automl.engine</i>), 189	<i>evalml.data_checks.class_imbalance_data_check</i> module, 212
EngineComputation (class in <i>evalml.automl.engine.engine_base</i>), 182	<i>evalml.data_checks.data_check</i> module, 213
EnsembleMissingPipelinesError, 265, 267	<i>evalml.data_checks.data_check_action</i> module, 214
<code>error()</code> (<i>evalml.automl.engine.engine_base.JobLogger</i> method), 183	<i>evalml.data_checks.data_check_action_code</i> module, 215
<code>error_contains_nan</code> (in <i>module evalml.data_checks.datetime_nan_data_check</i>), 223	<i>evalml.data_checks.data_check_message</i> module, 216
<code>error_contains_nan</code> (in <i>module evalml.data_checks.natural_language_nan_data_check</i>), 230	<i>evalml.data_checks.data_check_message_code</i> module, 217
Estimator (class in <i>evalml.pipelines</i>), 1161	<i>evalml.data_checks.data_check_message_type</i> module, 219
Estimator (class in <i>evalml.pipelines.components</i>), 967	<i>evalml.data_checks.data_checks</i> module, 220
Estimator (class in <i>evalml.pipelines.components.estimators</i>), 656	<i>evalml.data_checks.datetime_format_data_check</i> module, 220
Estimator (class in <i>evalml.pipelines.components.estimators.estimator</i>), 628	<i>evalml.data_checks.datetime_nan_data_check</i> module, 222
<i>evalml</i> module, 166	<i>evalml.data_checks.default_data_checks</i> module, 223
	<i>evalml.data_checks.highly_null_data_check</i> module, 224

[evalml.data_checks.id_columns_data_check](#)
 module, 226

[evalml.data_checks.invalid_targets_data_check](#)
 module, 227

[evalml.data_checks.multicollinearity_data_check](#)
 module, 228

[evalml.data_checks.natural_language_nan_data_check](#)
 module, 229

[evalml.data_checks.no_variance_data_check](#)
 module, 231

[evalml.data_checks.outliers_data_check](#)
 module, 232

[evalml.data_checks.sparsity_data_check](#)
 module, 233

[evalml.data_checks.target_distribution_data_check](#)
 module, 235

[evalml.data_checks.target_leakage_data_check](#)
 module, 236

[evalml.data_checks.uniqueness_data_check](#)
 module, 237

[evalml.data_checks.utils](#)
 module, 239

[evalml.demos](#)
 module, 260

[evalml.demos.breast_cancer](#)
 module, 260

[evalml.demos.churn](#)
 module, 261

[evalml.demos.diabetes](#)
 module, 261

[evalml.demos.fraud](#)
 module, 262

[evalml.demos.weather](#)
 module, 262

[evalml.demos.wine](#)
 module, 263

[evalml.exceptions](#)
 module, 264

[evalml.exceptions.exceptions](#)
 module, 264

[evalml.model_family](#)
 module, 269

[evalml.model_family.model_family](#)
 module, 269

[evalml.model_family.utils](#)
 module, 270

[evalml.model_understanding](#)
 module, 271

[evalml.model_understanding.force_plots](#)
 module, 278

[evalml.model_understanding.graphs](#)
 module, 279

[evalml.model_understanding.permutation_importance](#)
 module, 289

[evalml.model_understanding.prediction_explanations](#)
 module, 271

[evalml.model_understanding.prediction_explanations](#)
 module, 271

[evalml.objectives](#)
 module, 303

[evalml.objectives.binary_classification_objective](#)
 module, 303

[evalml.objectives.cost_benefit_matrix](#)
 module, 306

[evalml.objectives.fraud_cost](#)
 module, 309

[evalml.objectives.lead_scoring](#)
 module, 312

[evalml.objectives.multiclass_classification_objective](#)
 module, 314

[evalml.objectives.objective_base](#)
 module, 317

[evalml.objectives.regression_objective](#)
 module, 319

[evalml.objectives.sensitivity_low_alert](#)
 module, 321

[evalml.objectives.standard_metrics](#)
 module, 324

[evalml.objectives.time_series_regression_objective](#)
 module, 387

[evalml.objectives.utils](#)
 module, 389

[evalml.pipelines](#)
 module, 474

[evalml.pipelines.binary_classification_pipeline](#)
 module, 1064

[evalml.pipelines.binary_classification_pipeline_mixin](#)
 module, 1069

[evalml.pipelines.classification_pipeline](#)
 module, 1070

[evalml.pipelines.component_graph](#)
 module, 1075

[evalml.pipelines.components](#)
 module, 474

[evalml.pipelines.components.component_base](#)
 module, 918

[evalml.pipelines.components.component_base_meta](#)
 module, 920

[evalml.pipelines.components.ensemble](#)
 module, 474

[evalml.pipelines.components.ensemble.sklearn_stackable](#)
 module, 474

[evalml.pipelines.components.ensemble.sklearn_stackable](#)
 module, 477

[evalml.pipelines.components.ensemble.sklearn_stackable](#)
 module, 480

[evalml.pipelines.components.estimators](#)
 module, 491

`evalml.pipelines.components.estimators.categorical` `evalml.pipelines.components.transformers`
module, 491 module, 697

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.column_selector`
module, 491 module, 836

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.dimensionality_reduction`
module, 494 module, 697

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.dimensionality_reduction`
module, 497 module, 697

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.dimensionality_reduction`
module, 500 module, 700

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.encoders`
module, 503 module, 707

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.encoders`
module, 506 module, 707

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.encoders`
module, 509 module, 711

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.feature_selector`
module, 513 module, 720

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.feature_selector`
module, 516 module, 720

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.feature_selector`
module, 518 module, 723

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.feature_selector`
module, 521 module, 726

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 628 module, 738

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 555 module, 738

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 555 module, 741

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 559 module, 744

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 561 module, 746

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 564 module, 760

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 568 module, 760

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 570 module, 763

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 574 module, 766

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 577 module, 768

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 580 module, 771

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 583 module, 773

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 585 module, 776

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 588 module, 778

`evalml.pipelines.components.estimators.categorical_pipeline_base` `evalml.pipelines.components.transformers.imputers`
module, 591 module, 781

<code>evalml.pipelines.components.transformer_sampler</code>	<code>evalml.problem_types.transformer_sampler</code>
module, 783	module, 1274
<code>evalml.pipelines.components.transformer_sampler.transform_primitive_components</code>	<code>evalml.problem_types.transform_primitive_components</code>
module, 786	module, 1278
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 818	module, 1278
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 818	module, 1279
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 820	module, 1281
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 823	module, 1282
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 831	module, 1283
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 831	module, 1288
<code>evalml.pipelines.components.transformer_sampler.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>
module, 845	module, 1288
<code>evalml.pipelines.components.utils</code>	<code>evalml.utils.cli_utils</code>
module, 921	module, 1288
<code>evalml.pipelines.multiclass_classification_pipeline</code>	<code>evalml.pipelines.gen_utils</code>
module, 1079	module, 1289
<code>evalml.pipelines.pipeline_base</code>	<code>evalml.utils.logger</code>
module, 1084	module, 1293
<code>evalml.pipelines.pipeline_meta</code>	<code>evalml.utils.update_checker</code>
module, 1089	module, 1294
<code>evalml.pipelines.regression_pipeline</code>	<code>evalml.utils.woodwork_utils</code>
module, 1089	module, 1294
<code>evalml.pipelines.time_series_classification_pipeline</code>	<code>evalml.pipeline_base</code>
module, 1094	(in module <code>evalml.automl.engine</code>), 190
<code>evalml.pipelines.time_series_pipeline_base</code>	<code>evalml.pipeline_base</code>
module, 1112	(in module <code>evalml.automl.engine.engine_base</code>), 182
<code>evalml.pipelines.time_series_regression_pipeline</code>	<code>evalml.objectives.binary_classification_objective</code>
module, 1117	(property), 304
<code>evalml.pipelines.utils</code>	<code>evalml.objectives.BinaryClassificationObjective</code>
module, 1122	(property), 408
<code>evalml.preprocessing</code>	<code>evalml.objectives.multiclass_classification_objective</code>
module, 1261	(property), 315
<code>evalml.preprocessing.data_splitters</code>	<code>evalml.objectives.MulticlassClassificationObjective</code>
module, 1261	(property), 447
<code>evalml.preprocessing.data_splitters.balance_split</code>	<code>evalml.objectives.objective_base.ObjectiveBase</code>
module, 1261	(property), 318
<code>evalml.preprocessing.data_splitters.sample_split</code>	<code>evalml.objectives.ObjectiveBase</code>
module, 1263	(property), 449
<code>evalml.preprocessing.data_splitters.time_series_split</code>	<code>evalml.objectives.regression_objective.RegressionObjective</code>
module, 1263	(property), 320
<code>evalml.preprocessing.data_splitters.train_test_split</code>	<code>evalml.objectives.RegressionObjective</code>
module, 1265	(property), 467
<code>evalml.preprocessing.utils</code>	<code>evalml.objectives.time_series_regression_objective</code>
module, 1269	(property), 388
<code>evalml.problem_types</code>	<code>explain_predictions()</code>
module, 1273	(in module <code>evalml.model_understanding</code>), 294
<code>evalml.problem_types.problem_types</code>	<code>explain_predictions()</code>
module, 1273	(in module <code>evalml.model_understanding.prediction_explanations</code>),

276
[explain_predictions\(\)](#) (in module [F1Weighted](#) (class in [evalml.objectives](#)), 419
[evalml.model_understanding.prediction_explanations.explain_predictions\(\)](#) (class in [evalml.objectives.standard_metrics](#)), 347
272
[explain_predictions_best_worst\(\)](#) (in module [evalml.model_understanding](#)), 294
[explain_predictions_best_worst\(\)](#) (in module [evalml.model_understanding.prediction_explanations](#)), 277
[explain_predictions_best_worst\(\)](#) (in module [evalml.model_understanding.prediction_explanations.explain_predictions\(\)](#)), 273
[ExplainPredictionsStage](#) (class in [evalml.model_understanding.prediction_explanations.explain_predictions\(\)](#)), 275
[ExpVariance](#) (class in [evalml.objectives](#)), 412
[ExpVariance](#) (class in [evalml.objectives.standard_metrics](#)), 340
[ExtraTreesClassifier](#) (class in [evalml.pipelines](#)), 1164
[ExtraTreesClassifier](#) (class in [evalml.pipelines.components](#)), 970
[ExtraTreesClassifier](#) (class in [evalml.pipelines.components.estimators](#)), 658
[ExtraTreesClassifier](#) (class in [evalml.pipelines.components.estimators.classifiers](#)), 535
[ExtraTreesClassifier](#) (class in [evalml.pipelines.components.estimators.classifiers.et_classifier](#)), 503
[ExtraTreesRegressor](#) (class in [evalml.pipelines](#)), 1166
[ExtraTreesRegressor](#) (class in [evalml.pipelines.components](#)), 973
[ExtraTreesRegressor](#) (class in [evalml.pipelines.components.estimators](#)), 661
[ExtraTreesRegressor](#) (class in [evalml.pipelines.components.estimators.regressors](#)), 607
[ExtraTreesRegressor](#) (class in [evalml.pipelines.components.estimators.regressors.et_regressor](#)), 571
F
[F1](#) (class in [evalml.objectives](#)), 414
[F1](#) (class in [evalml.objectives.standard_metrics](#)), 341
[F1Macro](#) (class in [evalml.objectives](#)), 416
[F1Macro](#) (class in [evalml.objectives.standard_metrics](#)), 343
[F1Micro](#) (class in [evalml.objectives](#)), 417
[F1Micro](#) (class in [evalml.objectives.standard_metrics](#)), 345
[F1Weighted](#) (class in [evalml.objectives](#)), 419
[feature_importance\(\)](#) (class in [evalml.pipelines.ARIMARegressor](#) property), 1127
[feature_importance\(\)](#) (class in [evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) property), 1066
[feature_importance\(\)](#) (class in [evalml.pipelines.BinaryClassificationPipeline](#) property), 1130
[feature_importance\(\)](#) (class in [evalml.pipelines.CatBoostClassifier](#) property), 1135
[feature_importance\(\)](#) (class in [evalml.pipelines.CatBoostRegressor](#) property), 1137
[feature_importance\(\)](#) (class in [evalml.pipelines.classification_pipeline.ClassificationPipeline](#) property), 1072
[feature_importance\(\)](#) (class in [evalml.pipelines.ClassificationPipeline](#) property), 1140
[feature_importance\(\)](#) (class in [evalml.pipelines.components.ARIMARegressor](#) property), 929
[feature_importance\(\)](#) (class in [evalml.pipelines.components.BaselineClassifier](#) property), 931
[feature_importance\(\)](#) (class in [evalml.pipelines.components.BaselineRegressor](#) property), 933
[feature_importance\(\)](#) (class in [evalml.pipelines.components.CatBoostClassifier](#) property), 936
[feature_importance\(\)](#) (class in [evalml.pipelines.components.CatBoostRegressor](#) property), 938
[feature_importance\(\)](#) (class in [evalml.pipelines.components.DecisionTreeClassifier](#) property), 946
[feature_importance\(\)](#) (class in [evalml.pipelines.components.DecisionTreeRegressor](#) property), 949
[feature_importance\(\)](#) (class in [evalml.pipelines.components.ElasticNetClassifier](#) property), 962
[feature_importance\(\)](#) (class in [evalml.pipelines.components.ElasticNetRegressor](#) property), 965
[feature_importance\(\)](#) (class in [evalml.pipelines.components.ensemble.sklearn_stacked_ensemble](#) property), 476

feature_importance() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifierElasticNetClassifier), 479	feature_importance() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifierElasticNetClassifier), 535
feature_importance() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifierElasticNetClassifier), 482	feature_importance() (evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.SklearnStackedEnsembleClassifierElasticNetClassifier), 505
feature_importance() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier), 484	feature_importance() (evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier), 538
feature_importance() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier), 487	feature_importance() (evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier), 508
feature_importance() (evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier), 490	feature_importance() (evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier), 541
feature_importance() (evalml.pipelines.components.Estimator), 969	feature_importance() (evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier), 512
feature_importance() (evalml.pipelines.components.estimators.ARIMAREgressor), 634	feature_importance() (evalml.pipelines.components.estimators.classifiers.LightGBMClassifier), 544
feature_importance() (evalml.pipelines.components.estimators.BaselineClassifier), 636	feature_importance() (evalml.pipelines.components.estimators.classifiers.logistic_regression.LogisticRegression), 515
feature_importance() (evalml.pipelines.components.estimators.BaselineRegressor), 638	feature_importance() (evalml.pipelines.components.estimators.classifiers.LogisticRegression), 547
feature_importance() (evalml.pipelines.components.estimators.CatBoostClassifier), 641	feature_importance() (evalml.pipelines.components.estimators.classifiers.RandomForestClassifier), 549
feature_importance() (evalml.pipelines.components.estimators.CatBoostRegressor), 644	feature_importance() (evalml.pipelines.components.estimators.classifiers.rf_classifier.RandomForestClassifier), 517
feature_importance() (evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier), 493	feature_importance() (evalml.pipelines.components.estimators.classifiers.svm_classifier.SVMClassifier), 520
feature_importance() (evalml.pipelines.components.estimators.classifiers.BaselineClassifier), 526	feature_importance() (evalml.pipelines.components.estimators.classifiers.SVMClassifier), 552
feature_importance() (evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier), 496	feature_importance() (evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier), 523
feature_importance() (evalml.pipelines.components.estimators.classifiers.CatBoostClassifier), 529	feature_importance() (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier), 554
feature_importance() (evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier), 499	feature_importance() (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier), 647
feature_importance() (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier), 532	feature_importance() (evalml.pipelines.components.estimators.DecisionTreeRegressor), 650
feature_importance() (evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.ElasticNetClassifier), 502	feature_importance() (evalml.pipelines.components.estimators.ElasticNetClassifier), 653

<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.ElasticNetRegressor</code> <code>property</code>), 655	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.CatBoostRegressor</code> <code>property</code>), 601
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.Estimator</code> <code>property</code>), 657	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.decision_tree</code> <code>property</code>), 567
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.estimator.Estimator</code> <code>property</code>), 629	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor</code> <code>property</code>), 604
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.ExtraTreesClassifier</code> <code>property</code>), 660	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.elasticnet_regressor</code> <code>property</code>), 569
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.ExtraTreesRegressor</code> <code>property</code>), 663	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.ElasticNetRegressor</code> <code>property</code>), 606
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.KNeighborsClassifier</code> <code>property</code>), 666	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.et_regressor</code> <code>property</code>), 573
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.LightGBMClassifier</code> <code>property</code>), 669	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor</code> <code>property</code>), 609
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.LightGBMRegressor</code> <code>property</code>), 672	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.lightgbm_regressor</code> <code>property</code>), 576
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.LinearRegressor</code> <code>property</code>), 674	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.LightGBMRegressor</code> <code>property</code>), 612
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.LogisticRegressionClassifier</code> <code>property</code>), 677	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.linear_regressor</code> <code>property</code>), 579
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.ProphetRegressor</code> <code>property</code>), 679	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.LinearRegressor</code> <code>property</code>), 615
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.RandomForestClassifier</code> <code>property</code>), 682	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.prophet_regressor</code> <code>property</code>), 582
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.RandomForestRegressor</code> <code>property</code>), 684	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.ProphetRegressor</code> <code>property</code>), 617
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.arma_regressor</code> <code>property</code>), 558	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.RandomForestRegressor</code> <code>property</code>), 620
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.ARIMARegressor</code> <code>property</code>), 596	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.rf_regressor</code> <code>property</code>), 584
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.baseline_regressor</code> <code>property</code>), 560	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.svm_regressor</code> <code>property</code>), 587
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.BaselineRegressor</code> <code>property</code>), 598	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.SVMRegressor</code> <code>property</code>), 622
<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.catboost_regressor</code> <code>property</code>), 563	<code>feature_importance()</code> (<code>evalml.pipelines.components.estimators.regressors.time_series_baseline_regressor</code> <code>property</code>), 590

feature_importance() (evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator property), 624	feature_importance() (evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator property), 1033	
feature_importance() (evalml.pipelines.components.estimators.regressors.xgboost.XGBRegressor property), 593	feature_importance() (evalml.pipelines.components.estimators.regressors.xgboost.XGBRegressor property), 1035	
feature_importance() (evalml.pipelines.components.estimators.regressors.XGBRegressor property), 627	feature_importance() (evalml.pipelines.components.SVMClassifier property), 1040	
feature_importance() (evalml.pipelines.components.estimators.SVMClassifier property), 686	feature_importance() (evalml.pipelines.components.SVMRegressor property), 1042	
feature_importance() (evalml.pipelines.components.estimators.SVMRegressor property), 689	feature_importance() (evalml.pipelines.components.TimeSeriesBaselineEstimator property), 1051	
feature_importance() (evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator property), 691	feature_importance() (evalml.pipelines.components.XGBoostClassifier property), 1060	
feature_importance() (evalml.pipelines.components.estimators.XGBoostClassifier property), 694	feature_importance() (evalml.pipelines.components.XGBoostRegressor property), 1063	
feature_importance() (evalml.pipelines.components.estimators.XGBoostRegressor property), 696	feature_importance() (evalml.pipelines.DecisionTreeClassifier property), 1148	
feature_importance() (evalml.pipelines.components.ExtraTreesClassifier property), 972	feature_importance() (evalml.pipelines.DecisionTreeRegressor property), 1151	
feature_importance() (evalml.pipelines.components.ExtraTreesRegressor property), 975	feature_importance() (evalml.pipelines.ElasticNetClassifier property), 1158	
feature_importance() (evalml.pipelines.components.KNeighborsClassifier property), 982	feature_importance() (evalml.pipelines.ElasticNetRegressor property), 1160	
feature_importance() (evalml.pipelines.components.LightGBMClassifier property), 985	feature_importance() (evalml.pipelines.Estimator property), 1162	
feature_importance() (evalml.pipelines.components.LightGBMRegressor property), 988	feature_importance() (evalml.pipelines.ExtraTreesClassifier property), 1166	
feature_importance() (evalml.pipelines.components.LinearRegressor property), 992	feature_importance() (evalml.pipelines.ExtraTreesRegressor property), 1169	
feature_importance() (evalml.pipelines.components.LogisticRegressionClassifier property), 995	feature_importance() (evalml.pipelines.KNeighborsClassifier property), 1174	
feature_importance() (evalml.pipelines.components.ProphetRegressor property), 1014	feature_importance() (evalml.pipelines.LightGBMClassifier property), 1176	
feature_importance() (evalml.pipelines.components.RandomForestClassifier property), 1016	feature_importance() (evalml.pipelines.LightGBMRegressor property), 1179	
feature_importance() (evalml.pipelines.components.RandomForestRegressor property), 1018	feature_importance() (evalml.pipelines.LinearRegressor property), 1181	
	feature_importance()	

<code>(evalml.pipelines.LogisticRegressionClassifier property), 1184</code>	<code>(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline property), 1119</code>
<code>feature_importance()</code> <code>(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline property), 1081</code>	<code>feature_importance()</code> <code>(evalml.pipelines.TimeSeriesBinaryClassificationPipeline property), 1234</code>
<code>feature_importance()</code> <code>(evalml.pipelines.MulticlassClassificationPipeline property), 1187</code>	<code>feature_importance()</code> <code>(evalml.pipelines.TimeSeriesClassificationPipeline property), 1240</code>
<code>feature_importance()</code> <code>(evalml.pipelines.pipeline_base.PipelineBase property), 1086</code>	<code>feature_importance()</code> <code>(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property), 1246</code>
<code>feature_importance()</code> <code>(evalml.pipelines.PipelineBase property), 1197</code>	<code>feature_importance()</code> <code>(evalml.pipelines.TimeSeriesRegressionPipeline property), 1251</code>
<code>feature_importance()</code> <code>(evalml.pipelines.ProphetRegressor property), 1201</code>	<code>feature_importance()</code> <code>(evalml.pipelines.XGBoostClassifier property), 1258</code>
<code>feature_importance()</code> <code>(evalml.pipelines.RandomForestClassifier property), 1203</code>	<code>feature_importance()</code> <code>(evalml.pipelines.XGBoostRegressor property), 1260</code>
<code>feature_importance()</code> <code>(evalml.pipelines.RandomForestRegressor property), 1205</code>	<code>FeatureSelector (class in evalml.pipelines), 1169</code>
<code>feature_importance()</code> <code>(evalml.pipelines.regression_pipeline.RegressionPipeline property), 1091</code>	<code>FeatureSelector (class in evalml.pipelines.components), 975</code>
<code>feature_importance()</code> <code>(evalml.pipelines.RegressionPipeline property), 1208</code>	<code>FeatureSelector (class in evalml.pipelines.components.transformers), 866</code>
<code>feature_importance()</code> <code>(evalml.pipelines.SklearnStackedEnsembleClassifier property), 1220</code>	<code>FeatureSelector (class in evalml.pipelines.components.transformers.feature_selection), 730</code>
<code>feature_importance()</code> <code>(evalml.pipelines.SklearnStackedEnsembleRegressor property), 1222</code>	<code>FeatureSelector (class in evalml.pipelines.components.transformers.feature_selection.feature_selection), 721</code>
<code>feature_importance()</code> <code>(evalml.pipelines.SVMClassifier property), 1227</code>	<code>fit() (evalml.pipelines.ARIMAREgressor method), 1127</code>
<code>feature_importance()</code> <code>(evalml.pipelines.SVMRegressor property), 1229</code>	<code>fit() (evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline method), 1066</code>
<code>feature_importance()</code> <code>(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1097</code>	<code>fit() (evalml.pipelines.BinaryClassificationPipeline method), 1130</code>
<code>feature_importance()</code> <code>(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1103</code>	<code>fit() (evalml.pipelines.CatBoostClassifier method), 1135</code>
<code>feature_importance()</code> <code>(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1108</code>	<code>fit() (evalml.pipelines.CatBoostRegressor method), 1137</code>
<code>feature_importance()</code> <code>(evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase property), 1114</code>	<code>fit() (evalml.pipelines.ClassificationPipeline method), 1072</code>
<code>feature_importance()</code>	<code>fit() (evalml.pipelines.ClassificationPipeline method), 1140</code>
	<code>fit() (evalml.pipelines.component_graph.ComponentGraph method), 1077</code>
	<code>fit() (evalml.pipelines.components.CatBoostClassifier method), 1144</code>
	<code>fit() (evalml.pipelines.components.ARIMAREgressor method), 929</code>
	<code>fit() (evalml.pipelines.components.BaselineClassifier method), 931</code>

`fit()` (`evalml.pipelines.components.BaselineRegressor` method), 933

`fit()` (`evalml.pipelines.components.CatBoostClassifier` method), 936

`fit()` (`evalml.pipelines.components.CatBoostRegressor` method), 938

`fit()` (`evalml.pipelines.components.component_base.ComponentBase` method), 919

`fit()` (`evalml.pipelines.components.ComponentBase` method), 940

`fit()` (`evalml.pipelines.components.DateTimeFeaturizer` method), 943

`fit()` (`evalml.pipelines.components.DecisionTreeClassifier` method), 946

`fit()` (`evalml.pipelines.components.DecisionTreeRegressor` method), 949

`fit()` (`evalml.pipelines.components.DelayedFeatureTransformer` method), 951

`fit()` (`evalml.pipelines.components.DFSTransformer` method), 953

`fit()` (`evalml.pipelines.components.DropColumns` method), 956

`fit()` (`evalml.pipelines.components.DropNullColumns` method), 957

`fit()` (`evalml.pipelines.components.DropRowsTransformer` method), 960

`fit()` (`evalml.pipelines.components.ElasticNetClassifier` method), 962

`fit()` (`evalml.pipelines.components.ElasticNetRegressor` method), 965

`fit()` (`evalml.pipelines.components.EmailFeaturizer` method), 967

`fit()` (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier` method), 476

`fit()` (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier` method), 479

`fit()` (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier` method), 482

`fit()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier` method), 484

`fit()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier` method), 487

`fit()` (`evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor` method), 490

`fit()` (`evalml.pipelines.components.Estimator` method), 969

`fit()` (`evalml.pipelines.components.estimators.ARIMARegressor` method), 634

`fit()` (`evalml.pipelines.components.estimators.BaselineClassifier` method), 636

`fit()` (`evalml.pipelines.components.estimators.BaselineRegressor` method), 638

`fit()` (`evalml.pipelines.components.estimators.CatBoostClassifier` method), 641

`fit()` (`evalml.pipelines.components.estimators.CatBoostRegressor` method), 644

`fit()` (`evalml.pipelines.components.estimators.classifiers.baseline_classifier` method), 493

`fit()` (`evalml.pipelines.components.estimators.classifiers.BaselineClassifier` method), 526

`fit()` (`evalml.pipelines.components.estimators.classifiers.catboost_classifier` method), 496

`fit()` (`evalml.pipelines.components.estimators.classifiers.CatBoostClassifier` method), 529

`fit()` (`evalml.pipelines.components.estimators.classifiers.decision_tree_classifier` method), 499

`fit()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` method), 532

`fit()` (`evalml.pipelines.components.estimators.classifiers.elasticnet_classifier` method), 502

`fit()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier` method), 535

`fit()` (`evalml.pipelines.components.estimators.classifiers.et_classifier.ExtraTreesClassifier` method), 505

`fit()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier` method), 538

`fit()` (`evalml.pipelines.components.estimators.classifiers.kneighbors_classifier` method), 508

`fit()` (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier` method), 541

`fit()` (`evalml.pipelines.components.estimators.classifiers.lightgbm_classifier` method), 512

`fit()` (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier` method), 544

`fit()` (`evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier` method), 515

`fit()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` method), 547

`fit()` (`evalml.pipelines.components.estimators.classifiers.SklearnStackedEnsembleClassifier` method), 549

`fit()` (`evalml.pipelines.components.estimators.classifiers.SklearnStackedEnsembleClassifier` method), 517

`fit()` (`evalml.pipelines.components.estimators.classifiers.svm_classifier.SVMClassifier` method), 520

`fit()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` method), 552

`fit()` (`evalml.pipelines.components.estimators.classifiers.xgboost_classifier` method), 523

`fit()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` method), 554

`fit()` (`evalml.pipelines.components.estimators.DecisionTreeClassifier` method), 647

`fit()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 650

`fit()` (`evalml.pipelines.components.estimators.ElasticNetClassifier` method), 653

`fit()` (`evalml.pipelines.components.estimators.ElasticNetRegressor` method), 655

`fit()` (`evalml.pipelines.components.estimators.Estimator` `fit()` (`evalml.pipelines.components.estimators.regressors.LinearRegressor`
 method), 657 method), 615
`fit()` (`evalml.pipelines.components.estimators.estimator.Estimator` (`evalml.pipelines.components.estimators.regressors.prophet_regressor`
 method), 630 method), 582
`fit()` (`evalml.pipelines.components.estimators.ExtraTreesClassifier` (`evalml.pipelines.components.estimators.regressors.ProphetRegressor`
 method), 660 method), 617
`fit()` (`evalml.pipelines.components.estimators.ExtraTreesRegressor` (`evalml.pipelines.components.estimators.regressors.RandomForestRegressor`
 method), 663 method), 620
`fit()` (`evalml.pipelines.components.estimators.KNeighborsClassifier` (`evalml.pipelines.components.estimators.regressors.rf_regressor.RandomForestRegressor`
 method), 666 method), 584
`fit()` (`evalml.pipelines.components.estimators.LightGBMClassifier` (`evalml.pipelines.components.estimators.regressors.svm_regressor.SVMRegressor`
 method), 669 method), 587
`fit()` (`evalml.pipelines.components.estimators.LightGBMRegressor` (`evalml.pipelines.components.estimators.regressors.SVMRegressor`
 method), 672 method), 622
`fit()` (`evalml.pipelines.components.estimators.LinearRegressor` (`evalml.pipelines.components.estimators.regressors.time_series_baseline_estimator.TimeSeriesBaselineEstimator`
 method), 674 method), 590
`fit()` (`evalml.pipelines.components.estimators.LogisticRegressionClassifier` (`evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator`
 method), 677 method), 624
`fit()` (`evalml.pipelines.components.estimators.ProphetRegressor` (`evalml.pipelines.components.estimators.regressors.xgboost_regressor.XGBoostRegressor`
 method), 679 method), 593
`fit()` (`evalml.pipelines.components.estimators.RandomForestClassifier` (`evalml.pipelines.components.estimators.regressors.XGBoostRegressor`
 method), 682 method), 627
`fit()` (`evalml.pipelines.components.estimators.RandomForestRegressor` (`evalml.pipelines.components.estimators.SVMClassifier`
 method), 684 method), 686
`fit()` (`evalml.pipelines.components.estimators.regressors.ann_regressor.ANNRegressor` (`evalml.pipelines.components.estimators.SVMRegressor`
 method), 558 method), 689
`fit()` (`evalml.pipelines.components.estimators.regressors.ARIMARegressor` (`evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator`
 method), 596 method), 691
`fit()` (`evalml.pipelines.components.estimators.regressors.baseline_estimator.BaselineEstimator` (`evalml.pipelines.components.estimators.XGBoostClassifier`
 method), 560 method), 694
`fit()` (`evalml.pipelines.components.estimators.regressors.BaselineRegressor` (`evalml.pipelines.components.estimators.XGBoostRegressor`
 method), 598 method), 696
`fit()` (`evalml.pipelines.components.estimators.regressors.BoostingRegressor` (`evalml.pipelines.components.estimators.ExtraTreesClassifier`
 method), 563 method), 972
`fit()` (`evalml.pipelines.components.estimators.regressors.CutBoostRegressor` (`evalml.pipelines.components.ExtraTreesRegressor`
 method), 601 method), 975
`fit()` (`evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor` (`evalml.pipelines.components.FeatureSelector`
 method), 567 method), 977
`fit()` (`evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor` (`evalml.pipelines.components.Imputer` method),
 method), 604 979
`fit()` (`evalml.pipelines.components.estimators.regressors.ElasticNetRegressor` (`evalml.pipelines.components.ElasticNetRegressor`
 method), 569 method), 982
`fit()` (`evalml.pipelines.components.estimators.regressors.ElasticNetRegressor` (`evalml.pipelines.components.LightGBMClassifier`
 method), 606 method), 985
`fit()` (`evalml.pipelines.components.estimators.regressors.EnRegressor` (`evalml.pipelines.components.LightGBMRegressor`
 method), 573 method), 988
`fit()` (`evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor` (`evalml.pipelines.components.LinearDiscriminantAnalysis`
 method), 609 method), 990
`fit()` (`evalml.pipelines.components.estimators.regressors.LightGBMRegressor` (`evalml.pipelines.components.LinearRegressor`
 method), 576 method), 992
`fit()` (`evalml.pipelines.components.estimators.regressors.LightGBMRegressor` (`evalml.pipelines.components.LogisticRegressionClassifier`
 method), 612 method), 995
`fit()` (`evalml.pipelines.components.estimators.regressors.LinReg` (`evalml.pipelines.components.LogTransformer`
 method), 579 method), 997

[fit \(\) \(evalml.pipelines.components.LSA method\), 999](#)
[fit \(\) \(evalml.pipelines.components.OneHotEncoder method\), 1002](#)
[fit \(\) \(evalml.pipelines.components.Oversampler method\), 1004](#)
[fit \(\) \(evalml.pipelines.components.PCA method\), 1007](#)
[fit \(\) \(evalml.pipelines.components.PerColumnImputer method\), 1009](#)
[fit \(\) \(evalml.pipelines.components.PolynomialDetrender method\), 1011](#)
[fit \(\) \(evalml.pipelines.components.ProphetRegressor method\), 1014](#)
[fit \(\) \(evalml.pipelines.components.RandomForestClassifier method\), 1016](#)
[fit \(\) \(evalml.pipelines.components.RandomForestRegressor method\), 1018](#)
[fit \(\) \(evalml.pipelines.components.RFClassifierSelectFromModel method\), 1021](#)
[fit \(\) \(evalml.pipelines.components.RFRegressorSelectFromModel method\), 1023](#)
[fit \(\) \(evalml.pipelines.components.SelectByType method\), 1026](#)
[fit \(\) \(evalml.pipelines.components.SelectColumns method\), 1028](#)
[fit \(\) \(evalml.pipelines.components.SimpleImputer method\), 1030](#)
[fit \(\) \(evalml.pipelines.components.SklearnStackedEnsembleClassifier method\), 1033](#)
[fit \(\) \(evalml.pipelines.components.SklearnStackedEnsembleRegressor method\), 1035](#)
[fit \(\) \(evalml.pipelines.components.StandardScaler method\), 1037](#)
[fit \(\) \(evalml.pipelines.components.SVMClassifier method\), 1040](#)
[fit \(\) \(evalml.pipelines.components.SVMRegressor method\), 1042](#)
[fit \(\) \(evalml.pipelines.components.TargetEncoder method\), 1044](#)
[fit \(\) \(evalml.pipelines.components.TargetImputer method\), 1046](#)
[fit \(\) \(evalml.pipelines.components.TextFeaturizer method\), 1049](#)
[fit \(\) \(evalml.pipelines.components.TimeSeriesBaselineEstimator method\), 1051](#)
[fit \(\) \(evalml.pipelines.components.Transformer method\), 1053](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selector.ColumnSelector method\), 837](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selector.DropColumns method\), 840](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selector.SelectByType method\), 842](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selector.SelectColumns method\), 844](#)
[fit \(\) \(evalml.pipelines.components.transformers.DateTimeFeaturizer method\), 852](#)
[fit \(\) \(evalml.pipelines.components.transformers.DelayedFeatureTransformer method\), 855](#)
[fit \(\) \(evalml.pipelines.components.transformers.DFSTransformer method\), 857](#)
[fit \(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityTransformer method\), 699](#)
[fit \(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityTransformer method\), 704](#)
[fit \(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityTransformer method\), 706](#)
[fit \(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityTransformer method\), 701](#)
[fit \(\) \(evalml.pipelines.components.transformers.DropColumns method\), 859](#)
[fit \(\) \(evalml.pipelines.components.transformers.DropNullColumns method\), 861](#)
[fit \(\) \(evalml.pipelines.components.transformers.DropRowsTransformer method\), 863](#)
[fit \(\) \(evalml.pipelines.components.transformers.EmailFeaturizer method\), 865](#)
[fit \(\) \(evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder method\), 709](#)
[fit \(\) \(evalml.pipelines.components.transformers.encoders.OneHotEncoder method\), 716](#)
[fit \(\) \(evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder method\), 713](#)
[fit \(\) \(evalml.pipelines.components.transformers.encoders.TargetEncoder method\), 719](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.feature_selector.FeatureSelector method\), 722](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.FeatureSelector method\), 731](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.rf_classifier.RFClassifier method\), 725](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.rf_classifier.RFClassifier method\), 728](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.RFClassifier method\), 734](#)
[fit \(\) \(evalml.pipelines.components.transformers.feature_selection.RFClassifier method\), 737](#)
[fit \(\) \(evalml.pipelines.components.transformers.FeatureSelector method\), 868](#)
[fit \(\) \(evalml.pipelines.components.transformers.Imputer method\), 870](#)
[fit \(\) \(evalml.pipelines.components.transformers.imputers.Imputer method\), 751](#)
[fit \(\) \(evalml.pipelines.components.transformers.imputers.imputer.Imputer method\), 740](#)
[fit \(\) \(evalml.pipelines.components.transformers.imputers.per_column_imputer.PerColumnImputer method\), 742](#)
[fit \(\) \(evalml.pipelines.components.transformers.imputers.PerColumnImputer method\), 742](#)

method), 754
fit () (evalml.pipelines.components.transformers.imputers.SimpleImputer (evalml.pipelines.components.transformers.preprocessing.polynomial
method), 745 method), 779
fit () (evalml.pipelines.components.transformers.imputers.SimpleImputer (evalml.pipelines.components.transformers.preprocessing.Polynomial
method), 756 method), 810
fit () (evalml.pipelines.components.transformers.imputers.TargetEncoder (evalml.pipelines.components.transformers.preprocessing.text_featu
method), 748 method), 782
fit () (evalml.pipelines.components.transformers.imputers.TargetEncoder (evalml.pipelines.components.transformers.preprocessing.text_tran
method), 758 method), 785
fit () (evalml.pipelines.components.transformers.LinearDiscriminantAnalysis (evalml.pipelines.components.transformers.preprocessing.TextFeatu
method), 872 method), 812
fit () (evalml.pipelines.components.transformers.LogTransformer (evalml.pipelines.components.transformers.preprocessing.TextTrans
method), 874 method), 814
fit () (evalml.pipelines.components.transformers.LSA fit () (evalml.pipelines.components.transformers.preprocessing.transform
method), 877 method), 787
fit () (evalml.pipelines.components.transformers.OneHotEncoder (evalml.pipelines.components.transformers.preprocessing.transform
method), 879 method), 789
fit () (evalml.pipelines.components.transformers.Oversampler fit () (evalml.pipelines.components.transformers.preprocessing.URLFeat
method), 882 method), 816
fit () (evalml.pipelines.components.transformers.PCA fit () (evalml.pipelines.components.transformers.RFClassifierSelectFrom
method), 884 method), 892
fit () (evalml.pipelines.components.transformers.PerColumnImputer (evalml.pipelines.components.transformers.RFRegressorSelectFrom
method), 887 method), 894
fit () (evalml.pipelines.components.transformers.PolynomialDetector (evalml.pipelines.components.transformers.samplers.base_sampler.
method), 889 method), 819
fit () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer (evalml.pipelines.components.transformers.samplers.Oversampler
method), 761 method), 828
fit () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer (evalml.pipelines.components.transformers.samplers.oversampler.C
method), 792 method), 822
fit () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer (evalml.pipelines.components.transformers.samplers.undersampler.
method), 764 method), 830
fit () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer (evalml.pipelines.components.transformers.samplers.undersampler.
method), 795 method), 825
fit () (evalml.pipelines.components.transformers.preprocessing.DesStandardScaler (evalml.pipelines.components.transformers.scalers.standard_scaler.
method), 797 method), 833
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.scalers.StandardScaler
method), 767 method), 835
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.scalers.StandardScaler
method), 769 method), 897
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.SelectColumns
method), 799 method), 899
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.SimpleImputer
method), 801 method), 901
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.StandardScaler
method), 803 method), 903
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.TargetEncoder
method), 772 method), 905
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.TargetImputer
method), 774 method), 908
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.TextFeaturizer
method), 805 method), 910
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.Transformer
method), 808 method), 912
fit () (evalml.pipelines.components.transformers.preprocessing.DropNullColumns (evalml.pipelines.components.transformers.transformer.TargetTrans

method), 846

fit () (evalml.pipelines.components.transformers.transformer_base.TransformerBase method), 848

fit () (evalml.pipelines.components.transformers.Undersampler method), 915

fit () (evalml.pipelines.components.transformers.URLFeaturizer method), 917

fit () (evalml.pipelines.components.Undersampler method), 1056

fit () (evalml.pipelines.components.URLFeaturizer method), 1058

fit () (evalml.pipelines.components.utils.WrappedSKClassifier method), 922

fit () (evalml.pipelines.components.utils.WrappedSKRegressor method), 924

fit () (evalml.pipelines.components.XGBoostClassifier method), 1060

fit () (evalml.pipelines.components.XGBoostRegressor method), 1063

fit () (evalml.pipelines.DecisionTreeClassifier method), 1148

fit () (evalml.pipelines.DecisionTreeRegressor method), 1151

fit () (evalml.pipelines.DelayedFeatureTransformer method), 1153

fit () (evalml.pipelines.DFSSTransformer method), 1155

fit () (evalml.pipelines.ElasticNetClassifier method), 1158

fit () (evalml.pipelines.ElasticNetRegressor method), 1160

fit () (evalml.pipelines.Estimator method), 1163

fit () (evalml.pipelines.ExtraTreesClassifier method), 1166

fit () (evalml.pipelines.ExtraTreesRegressor method), 1169

fit () (evalml.pipelines.FeatureSelector method), 1171

fit () (evalml.pipelines.KNeighborsClassifier method), 1174

fit () (evalml.pipelines.LightGBMClassifier method), 1176

fit () (evalml.pipelines.LightGBMRegressor method), 1179

fit () (evalml.pipelines.LinearRegressor method), 1181

fit () (evalml.pipelines.LogisticRegressionClassifier method), 1184

fit () (evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method), 1081

fit () (evalml.pipelines.MulticlassClassificationPipeline method), 1187

fit () (evalml.pipelines.OneHotEncoder method), 1191

fit () (evalml.pipelines.PerColumnImputer method), 1194

fit () (evalml.pipelines.pipeline_base.PipelineBase method), 1086

fit () (evalml.pipelines.PipelineBase method), 1197

fit () (evalml.pipelines.ProphetRegressor method), 1201

fit () (evalml.pipelines.RandomForestClassifier method), 1203

fit () (evalml.pipelines.RandomForestRegressor method), 1205

fit () (evalml.pipelines.regression_pipeline.RegressionPipeline method), 1092

fit () (evalml.pipelines.RegressionPipeline method), 1208

fit () (evalml.pipelines.RFClassifierSelectFromModel method), 1212

fit () (evalml.pipelines.RFRegressorSelectFromModel method), 1215

fit () (evalml.pipelines.SimpleImputer method), 1217

fit () (evalml.pipelines.SklearnStackedEnsembleClassifier method), 1220

fit () (evalml.pipelines.SklearnStackedEnsembleRegressor method), 1222

fit () (evalml.pipelines.StandardScaler method), 1224

fit () (evalml.pipelines.SVMClassifier method), 1227

fit () (evalml.pipelines.SVMRegressor method), 1229

fit () (evalml.pipelines.TargetEncoder method), 1231

fit () (evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline method), 1097

fit () (evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method), 1103

fit () (evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline method), 1108

fit () (evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase method), 1114

fit () (evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method), 1119

fit () (evalml.pipelines.TimeSeriesBinaryClassificationPipeline method), 1234

fit () (evalml.pipelines.TimeSeriesClassificationPipeline method), 1240

fit () (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method), 1246

fit () (evalml.pipelines.TimeSeriesRegressionPipeline method), 1251

fit () (evalml.pipelines.Transformer method), 1255

fit () (evalml.pipelines.XGBoostClassifier method), 1258

fit () (evalml.pipelines.XGBoostRegressor method), 1260

fit_features () (evalml.pipelines.component_graph.ComponentGraph method), 1077

fit_features () (evalml.pipelines.ComponentGraph method), 1144

fit_resample () (evalml.preprocessing.data_splitters.balanced_classification_splitter.BalancedClassifierSplitter method), 1262

`fit_resample()` (`evalml.preprocessing.data_splitters.BalancedClassificationSampler` method), 1266
`fit_resample()` (`evalml.preprocessing.data_splitters.sampler_base.SimpleBaseSampler` method), 1263
`fit_resample()` (`evalml.preprocessing.data_splitters.SimpleBaseSampler` method), 1267
`fit_transform()` (`evalml.pipelines.components.DateTimeFeaturizer` method), 943
`fit_transform()` (`evalml.pipelines.components.DelayedFeatureTransformer` method), 951
`fit_transform()` (`evalml.pipelines.components.DFTransformer` method), 954
`fit_transform()` (`evalml.pipelines.components.DropColumns` method), 956
`fit_transform()` (`evalml.pipelines.components.DropNullColumns` method), 958
`fit_transform()` (`evalml.pipelines.components.DropRowsTransformer` method), 960
`fit_transform()` (`evalml.pipelines.components.EmailFeaturizer` method), 967
`fit_transform()` (`evalml.pipelines.components.FeatureSelector` method), 977
`fit_transform()` (`evalml.pipelines.components.Imputer` method), 979
`fit_transform()` (`evalml.pipelines.components.LinearDiscriminantAnalysis` method), 990
`fit_transform()` (`evalml.pipelines.components.LogTransformer` method), 997
`fit_transform()` (`evalml.pipelines.components.LSA` method), 999
`fit_transform()` (`evalml.pipelines.components.OneHotEncoder` method), 1002
`fit_transform()` (`evalml.pipelines.components.Oversampler` method), 1005
`fit_transform()` (`evalml.pipelines.components.PCA` method), 1007
`fit_transform()` (`evalml.pipelines.components.PerColumnImputer` method), 1009
`fit_transform()` (`evalml.pipelines.components.PolynomialDegreeReducer` method), 1011
`fit_transform()` (`evalml.pipelines.components.RFClassifierSelectFromModel` method), 1021
`fit_transform()` (`evalml.pipelines.components.RFRegressorSelectFromModel` method), 1024
`fit_transform()` (`evalml.pipelines.components.SelectByType` method), 1026
`fit_transform()` (`evalml.pipelines.components.SelectColumns` method), 1028
`fit_transform()` (`evalml.pipelines.components.SimpleImputer` method), 1030
`fit_transform()` (`evalml.pipelines.components.StandardScaler` method), 1037
`fit_transform()` (`evalml.pipelines.components.TargetEncoder` method), 1044

Index	1403
--------------	-------------

`fit_transform()` (*evalml.pipelines.components.transformers.SimpleImputer* method), 901
`fit_transform()` (*evalml.pipelines.components.transformers.StandardScaler* method), 903
`fit_transform()` (*evalml.pipelines.components.transformers.TargetEncoder* method), 905
`fit_transform()` (*evalml.pipelines.components.transformers.TargetEncoderPipeline* method), 908
`fit_transform()` (*evalml.pipelines.components.transformers.TextFeaturizer* method), 910
`fit_transform()` (*evalml.pipelines.components.transformers.Transformer* method), 912
`fit_transform()` (*evalml.pipelines.components.transformers.transformer_pipeline_base.TransformersPipeline* method), 846
`fit_transform()` (*evalml.pipelines.components.transformers.transformer_pipeline_base.TransformersPipeline* method), 849
`fit_transform()` (*evalml.pipelines.components.transformers.Undersampler* method), 915
`fit_transform()` (*evalml.pipelines.components.transformers.URLFeaturizer* method), 917
`fit_transform()` (*evalml.pipelines.components.Undersampler* method), 1056
`fit_transform()` (*evalml.pipelines.components.URLFeaturizer* method), 1130
`fit_transform()` (*evalml.pipelines.DelayedFeatureTransformer* method), 1073
`fit_transform()` (*evalml.pipelines.DFSTransformer* method), 1155
`fit_transform()` (*evalml.pipelines.FeatureSelector* method), 1171
`fit_transform()` (*evalml.pipelines.OneHotEncoder* method), 1191
`fit_transform()` (*evalml.pipelines.PerColumnImputer* method), 1194
`fit_transform()` (*evalml.pipelines.RFClassifierSelectFromModel* method), 1212
`fit_transform()` (*evalml.pipelines.RFRegressorSelectFromModel* method), 1215
`fit_transform()` (*evalml.pipelines.SimpleImputer* method), 1217
`fit_transform()` (*evalml.pipelines.StandardScaler* method), 1224
`fit_transform()` (*evalml.pipelines.TargetEncoder* method), 1231
`fit_transform()` (*evalml.pipelines.Transformer* method), 1255
`force_plot()` (in module *evalml.model_understanding.force_plots*), 278
FraudCost (class in *evalml.objectives*), 421
FraudCost (class in *evalml.objectives.fraud_cost*), 309
`full_rankings()` (*evalml.automl.automl_search.AutoMLSearch* property), 195
`full_rankings()` (*evalml.automl.AutoMLSearch* property), 207
`full_rankings()` (*evalml.AutoMLSearch* property), 202
G
`generate_component_code()` (in module *evalml.pipelines.components.utils*), 921
`generate_order()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1077
`generate_order()` (*evalml.pipelines.ComponentGraph* method), 1144
`generate_pipeline_code()` (in module *evalml.pipelines.components.utils*), 928
`get_all_objective_names()` (in module *evalml.pipelines.components.utils*), 390
`get_all_objective_names()` (in module *evalml.pipelines.components.utils*), 390
`get_best_sampler_for_data()` (in module *evalml.pipelines.components.utils*), 201
`get_component()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1066
`get_component()` (*evalml.pipelines.BinaryClassificationPipeline* method), 1066
`get_component()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1130
`get_component()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073
`get_component()` (*evalml.pipelines.ClassificationPipeline* method), 1140
`get_component()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1077
`get_component()` (*evalml.pipelines.ComponentGraph* method), 1145
`get_component()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1081
`get_component()` (*evalml.pipelines.MulticlassClassificationPipeline* method), 1187
`get_component()` (*evalml.pipelines.pipeline_base.PipelineBase* method), 1086
`get_component()` (*evalml.pipelines.PipelineBase* method), 1197
`get_component()` (*evalml.pipelines.regression_pipeline.RegressionPipeline* method), 1092
`get_component()` (*evalml.pipelines.RegressionPipeline* method), 1208
`get_component()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1097
`get_component()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1103
`get_component()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* method), 1109
`get_component()` (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase* method), 1114
`get_component()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* method), 1120

`get_component()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* method), 1235
`get_component()` (*evalml.pipelines.TimeSeriesClassificationPipeline* method), 1240
`get_component()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1246
`get_component()` (*evalml.pipelines.TimeSeriesRegressionPipeline* method), 1251
`get_core_objective_names()` (*in module evalml.objectives*), 423
`get_core_objective_names()` (*in module evalml.objectives.utils*), 390
`get_core_objectives()` (*in module evalml.objectives*), 423
`get_core_objectives()` (*in module evalml.objectives.utils*), 390
`get_default_primary_search_objective()` (*in module evalml.automl*), 209
`get_default_primary_search_objective()` (*in module evalml.automl.utils*), 201
`get_estimators()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1077
`get_estimators()` (*evalml.pipelines.ComponentGraph* method), 1145
`get_estimators()` (*in module evalml.pipelines.components.utils*), 921
`get_evalml_root()` (*in module evalml.utils.cli_utils*), 1289
`get_feature_names()` (*evalml.pipelines.components.DateTimeFeaturizer* method), 943
`get_feature_names()` (*evalml.pipelines.components.OneHotEncoder* method), 1002
`get_feature_names()` (*evalml.pipelines.components.TargetEncoder* method), 1045
`get_feature_names()` (*evalml.pipelines.components.transformers.DateTimeFeaturizer* method), 853
`get_feature_names()` (*evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder* method), 710
`get_feature_names()` (*evalml.pipelines.components.transformers.encoders.OneHotEncoder* method), 717
`get_feature_names()` (*evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder* method), 713
`get_feature_names()` (*evalml.pipelines.components.transformers.encoders.TargetEncoder* method), 720
`get_feature_names()` (*evalml.pipelines.components.transformers.OneHotEncoder* method), 880
`get_feature_names()` (*evalml.pipelines.components.transformers.preprocessing.datetime_featurizer.DateTimeFeaturizer* method), 762
`get_feature_names()` (*evalml.pipelines.OneHotEncoder* method), 1192
`get_feature_names()` (*evalml.pipelines.TargetEncoder* method), 1231
`get_hyperparameter_ranges()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1066
`get_hyperparameter_ranges()` (*evalml.pipelines.BinaryClassificationPipeline* method), 1130
`get_hyperparameter_ranges()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073
`get_hyperparameter_ranges()` (*evalml.pipelines.ClassificationPipeline* method), 1140
`get_hyperparameter_ranges()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1081
`get_hyperparameter_ranges()` (*evalml.pipelines.MulticlassClassificationPipeline* method), 1187
`get_hyperparameter_ranges()` (*evalml.pipelines.pipeline_base.PipelineBase* method), 1086
`get_hyperparameter_ranges()` (*evalml.pipelines.PipelineBase* method), 1197
`get_hyperparameter_ranges()` (*evalml.pipelines.RegressionPipeline* method), 1092
`get_hyperparameter_ranges()` (*evalml.pipelines.RegressionPipeline* method), 1208
`get_hyperparameter_ranges()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1097
`get_hyperparameter_ranges()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1103

`method`), 1109
`get_hyperparameter_ranges()` (`evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase` `method`), 1114
`get_hyperparameter_ranges()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` `method`), 1120
`get_hyperparameter_ranges()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` `method`), 1235
`get_hyperparameter_ranges()` (`evalml.pipelines.TimeSeriesClassificationPipeline` `method`), 1241
`get_hyperparameter_ranges()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` `method`), 1246
`get_hyperparameter_ranges()` (`evalml.pipelines.TimeSeriesRegressionPipeline` `method`), 1252
`get_importable_subclasses()` (in module `evalml.utils`), 1296
`get_importable_subclasses()` (in module `evalml.utils.gen_utils`), 1291
`get_inputs()` (`evalml.pipelines.component_graph.ComponentGraph` `method`), 1077
`get_inputs()` (`evalml.pipelines.ComponentGraph` `method`), 1145
`get_installed_packages()` (in module `evalml.utils.cli_utils`), 1289
`get_last_component()` (`evalml.pipelines.component_graph.ComponentGraph` `method`), 1078
`get_last_component()` (`evalml.pipelines.ComponentGraph` `method`), 1145
`get_linear_coefficients()` (in module `evalml.model_understanding`), 295
`get_linear_coefficients()` (in module `evalml.model_understanding.graphs`), 281
`get_logger()` (in module `evalml.utils`), 1297
`get_logger()` (in module `evalml.utils.logger`), 1293
`get_n_splits()` (`evalml.preprocessing.data_splitters.time_series_split.TimeSeriesSplit` `method`), 1264
`get_n_splits()` (`evalml.preprocessing.data_splitters.TimeSeriesSplit` `method`), 1267
`get_n_splits()` (`evalml.preprocessing.data_splitters.training_validation_split.TrainingValidationSplit` `static method`), 1265
`get_n_splits()` (`evalml.preprocessing.data_splitters.TrainingValidationSplit` `static method`), 1268
`get_n_splits()` (`evalml.preprocessing.TimeSeriesSplit` `method`), 1272
`get_n_splits()` (`evalml.preprocessing.TrainingValidationSplit` `static method`), 1272
`get_names()` (`evalml.pipelines.components.FeatureSelector` `method`), 977
`get_names()` (`evalml.pipelines.components.RFClassifierSelectFromModel` `method`), 1021
`get_names()` (`evalml.pipelines.components.RFRegressorSelectFromModel` `method`), 1024
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 722
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 731
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 725
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 729
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 734
`get_names()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` `method`), 737
`get_names()` (`evalml.pipelines.components.transformers.FeatureSelector` `method`), 868
`get_names()` (`evalml.pipelines.components.transformers.RFClassifierSelectFromModel` `method`), 892
`get_names()` (`evalml.pipelines.components.transformers.RFRegressorSelectFromModel` `method`), 895
`get_params()` (`evalml.pipelines.FeatureSelector` `method`), 1171
`get_names()` (`evalml.pipelines.RFClassifierSelectFromModel` `method`), 1212
`get_names()` (`evalml.pipelines.RFRegressorSelectFromModel` `method`), 1215
`get_non_core_objectives()` (in module `evalml.objectives`), 424
`get_non_core_objectives()` (in module `evalml.objectives.utils`), 390
`get_objective()` (in module `evalml.objectives`), 424
`get_objective()` (in module `evalml.objectives.utils`), 390
`get_params()` (`evalml.pipelines.components.estimators.ProphetRegressor` `method`), 679
`get_params()` (`evalml.pipelines.components.estimators.regressors.prophet.ProphetRegressor` `method`), 582
`get_params()` (`evalml.pipelines.components.estimators.regressors.ProphetRegressor` `method`), 1014
`get_params()` (`evalml.pipelines.components.ProphetRegressor` `method`), 1014
`get_params()` (`evalml.pipelines.components.utils.WrappedSKClassifier` `method`), 924
`get_params()` (`evalml.pipelines.components.utils.WrappedSKRegressor` `method`), 924
`get_params()` (`evalml.pipelines.ProphetRegressor` `method`), 1201
`get_pipeline()` (`evalml.automl.automl_search.AutoMLSearch` `method`), 195
`get_pipeline()` (`evalml.automl.AutoMLSearch` `method`), 207

`get_pipeline()` (*evalml.AutoMLSearch* method), 1303
`get_pipelines_from_component_graphs()` (in module *evalml.automl.utils*), 201
`get_prediction_vs_actual_data()` (in module *evalml.model_understanding*), 296
`get_prediction_vs_actual_data()` (in module *evalml.model_understanding.graphs*), 281
`get_prediction_vs_actual_over_time_data()` (in module *evalml.model_understanding*), 296
`get_prediction_vs_actual_over_time_data()` (in module *evalml.model_understanding.graphs*), 282
`get_random_seed()` (in module *evalml.utils*), 1297
`get_random_seed()` (in module *evalml.utils.gen_utils*), 1291
`get_random_state()` (in module *evalml.utils*), 1297
`get_random_state()` (in module *evalml.utils.gen_utils*), 1292
`get_result()` (*evalml.automl.engine.cf_engine.CFComputation* method), 177
`get_result()` (*evalml.automl.engine.dask_engine.DaskComputation* method), 179
`get_result()` (*evalml.automl.engine.engine_base.EngineComputation* method), 182
`get_result()` (*evalml.automl.engine.EngineComputation* method), 190
`get_result()` (*evalml.automl.engine.sequential_engine.SequentialComputation* method), 185
`get_sys_info()` (in module *evalml.utils.cli_utils*), 1289
Gini (class in *evalml.objectives*), 424
Gini (class in *evalml.objectives.standard_metrics*), 348
`graph()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1067
`graph()` (*evalml.pipelines.BinaryClassificationPipeline* method), 1130
`graph()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073
`graph()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073
`graph()` (*evalml.pipelines.ClassificationPipeline* method), 1141
`graph()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1078
`graph()` (*evalml.pipelines.ComponentGraph* method), 1145
`graph()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1081
`graph()` (*evalml.pipelines.MulticlassClassificationPipeline* method), 1187
`graph()` (*evalml.pipelines.pipeline_base.PipelineBase* method), 1087
`graph()` (*evalml.pipelines.PipelineBase* method), 1197
`graph()` (*evalml.pipelines.regression_pipeline.RegressionPipeline* method), 1092
`graph()` (*evalml.pipelines.RegressionPipeline* method), 1208
`graph()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1097
`graph()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1103
`graph()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1109
`graph()` (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase* method), 1115
`graph()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* method), 1120
`graph()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* method), 1235
`graph()` (*evalml.pipelines.TimeSeriesClassificationPipeline* method), 1241
`graph()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1246
`graph()` (*evalml.pipelines.TimeSeriesRegressionPipeline* method), 1252
`graph_binary_objective_vs_threshold()` (in module *evalml.model_understanding*), 296
`graph_binary_objective_vs_threshold()` (in module *evalml.model_understanding.graphs*), 282
`graph_confusion_matrix()` (in module *evalml.model_understanding*), 297
`graph_confusion_matrix()` (in module *evalml.model_understanding.graphs*), 282
`graph_feature_importance()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1067
`graph_feature_importance()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1130
`graph_feature_importance()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073
`graph_feature_importance()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1141
`graph_feature_importance()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1078
`graph_feature_importance()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1082
`graph_feature_importance()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1188
`graph_feature_importance()` (*evalml.pipelines.pipeline_base.PipelineBase* method), 1087
`graph_feature_importance()` (*evalml.pipelines.PipelineBase* method), 1197
`graph_feature_importance()` (*evalml.pipelines.PipelineBase* method), 1197
`graph_feature_importance()` (*evalml.pipelines.PipelineBase* method), 1197

(*evalml.pipelines.regression_pipeline.RegressionPipeline* (in module *evalml.model_understanding.graphs*),
 method), 1092 285
graph_feature_importance() (in module *evalml.model_understanding*), 299
 (*evalml.pipelines.RegressionPipeline* method), 1209 285
graph_feature_importance() (in module *evalml.model_understanding.graphs*), 285
 (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesBinaryClassificationPipeline* (in module
 method), 1097 *evalml.model_understanding*), 299
graph_feature_importance() (in module *evalml.model_understanding.graphs*), 285
 (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline*
 method), 1103 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.binary_classification_objective.BinaryClassificationObjective*
 (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline*
 method), 1109 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.BinaryClassificationObjective*
 (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase*), 408
 method), 1115 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective*
 (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline*
 method), 1120 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.MulticlassClassificationObjective*
 (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* property), 447
 method), 1235 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.objective_base.ObjectiveBase*
 (*evalml.pipelines.TimeSeriesClassificationPipeline* property), 318
 method), 1241 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.ObjectiveBase* property),
 (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* property), 449
 method), 1246 *greater_is_better()*
graph_feature_importance() (in module *evalml.objectives.regression_objective.RegressionObjective*
 (*evalml.pipelines.TimeSeriesRegressionPipeline* property), 320
 method), 1252 *greater_is_better()*
graph_force_plot() (in module *evalml.objectives.RegressionObjective* property), 467
evalml.model_understanding.force_plots), 279 *greater_is_better()*
graph_partial_dependence() (in module *evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective*
evalml.model_understanding), 297 property), 388
graph_partial_dependence() (in module *evalml.model_understanding.graphs*), 283
graph_permutation_importance() (in module *evalml.tuners.grid_search_tuner*), 1278
evalml.model_understanding), 298
graph_permutation_importance() (in module *evalml.model_understanding.graphs*), 284
graph_precision_recall_curve() (in module *evalml.model_understanding*), 298
graph_precision_recall_curve() (in module *evalml.model_understanding.graphs*), 284
graph_prediction_vs_actual() (in module *evalml.model_understanding*), 299
graph_prediction_vs_actual() (in module *evalml.model_understanding.graphs*), 284
graph_prediction_vs_actual_over_time() (in module *evalml.model_understanding*), 299
graph_prediction_vs_actual_over_time() (in module *evalml.model_understanding*), 299

H

handle_component_class() (in module *evalml.pipelines.components.utils*), 922
handle_model_family() (in module *evalml.model_family*), 270
handle_model_family() (in module *evalml.model_family.utils*), 270
handle_problem_types() (in module *evalml.problem_types*), 1276
handle_problem_types() (in module *evalml.problem_types.utils*), 1274
HighlyNullDataCheck (class in *evalml.data_checks*), 249

HighlyNullDataCheck (class in *evalml.data_checks.highly_null_data_check*), 224

I

IDColumnsDataCheck (class in *evalml.data_checks*), 251

IDColumnsDataCheck (class in *evalml.data_checks.id_columns_data_check*), 226

import_or_raise() (in module *evalml.utils*), 1297

import_or_raise() (in module *evalml.utils.gen_utils*), 1292

Imputer (class in *evalml.pipelines.components*), 978

Imputer (class in *evalml.pipelines.components.transformers*), 869

Imputer (class in *evalml.pipelines.components.transformers.imputers*), 750

Imputer (class in *evalml.pipelines.components.transformers.imputers.imputer*), 738

infer_feature_types() (in module *evalml.utils*), 1297

infer_feature_types() (in module *evalml.utils.woodwork_utils*), 1294

info() (*evalml.automl.engine.engine_base.JobLogger* method), 183

instantiate() (*evalml.pipelines.component_graph.ComponentGraph* method), 1078

instantiate() (*evalml.pipelines.ComponentGraph* method), 1145

InvalidTargetDataCheck (class in *evalml.data_checks*), 252

InvalidTargetDataCheck (class in *evalml.data_checks.invalid_targets_data_check*), 227

inverse_transform() (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 1067

inverse_transform() (*evalml.pipelines.BinaryClassificationPipeline* method), 1131

inverse_transform() (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 1073

inverse_transform() (*evalml.pipelines.ClassificationPipeline* method), 1141

inverse_transform() (*evalml.pipelines.component_graph.ComponentGraph* method), 1078

inverse_transform() (*evalml.pipelines.ComponentGraph* method), 1145

inverse_transform() (*evalml.pipelines.components.LogTransformer* method), 997

inverse_transform() (*evalml.pipelines.components.PolynomialDetrender* method), 1011

inverse_transform() (*evalml.pipelines.components.transformers.LogTransformer* method), 875

inverse_transform() (*evalml.pipelines.components.transformers.PolynomialDetrender* method), 889

inverse_transform() (*evalml.pipelines.components.transformers.preprocessing.log_transformer* method), 775

inverse_transform() (*evalml.pipelines.components.transformers.preprocessing.LogTransformer* method), 806

inverse_transform() (*evalml.pipelines.components.transformers.preprocessing.polynomial_detrender* method), 780

inverse_transform() (*evalml.pipelines.components.transformers.preprocessing.PolynomialDetrender* method), 810

inverse_transform() (*evalml.pipelines.components.transformers.transformer.TargetTransformer* method), 846

inverse_transform() (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* method), 1082

inverse_transform() (*evalml.pipelines.MulticlassClassificationPipeline* method), 1188

inverse_transform() (*evalml.pipelines.pipeline_base.PipelineBase* method), 1087

inverse_transform() (*evalml.pipelines.RegressionPipeline* method), 1197

inverse_transform() (*evalml.pipelines.regression_pipeline.RegressionPipeline* method), 1092

inverse_transform() (*evalml.pipelines.RegressionPipeline* method), 1209

inverse_transform() (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1097

inverse_transform() (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1104

inverse_transform() (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* method), 1109

inverse_transform()

`(evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase` (in module
`method), 1115` `evalml.problem_types), 1276`
`inverse_transform()` `is_classification()` (in module
`(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` `utils), 1275`
`method), 1120` `is_closed()` (`evalml.automl.engine.cf_engine.CFClient`
`inverse_transform()` `property), 176`
`(evalml.pipelines.TimeSeriesBinaryClassificationPipeline` `is_closed()` (`evalml.automl.engine.cf_engine.CFEngine`
`method), 1235` `property), 178`
`inverse_transform()` `is_closed()` (`evalml.automl.engine.CFEngine` `prop-`
`(evalml.pipelines.TimeSeriesClassificationPipeline` `erty), 186`
`method), 1241` `is_closed()` (`evalml.automl.engine.dask_engine.DaskEngine`
`inverse_transform()` `property), 180`
`(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` `is_defined()` (`evalml.automl.engine.DaskEngine`
`method), 1247` `property), 188`
`inverse_transform()` `is_defined_for_problem_type()`
`(evalml.pipelines.TimeSeriesRegressionPipeline` `(evalml.objectives.AccuracyBinary` `class`
`method), 1252` `method), 393`
`is_all_numeric()` (in module `evalml.utils), 1297` `is_defined_for_problem_type()`
`is_all_numeric()` (in module `evalml.utils.gen_utils), 1292` `(evalml.objectives.AccuracyMulticlass` `class`
`method), 395`
`is_binary()` (in module `evalml.problem_types), 1276` `is_defined_for_problem_type()`
`is_binary()` (in module `evalml.problem_types.utils), 1275` `(evalml.objectives.AUC` `class method), 397`
`is_bounded_like_percentage()` `(evalml.objectives.AUCMacro` `class method),`
`(evalml.objectives.binary_classification_objective.BinaryClassificationObjective`
`property), 305` `is_defined_for_problem_type()`
`is_bounded_like_percentage()` `(evalml.objectives.AUCMicro` `class method),`
`(evalml.objectives.BinaryClassificationObjective` `401`
`property), 408` `is_defined_for_problem_type()`
`is_bounded_like_percentage()` `(evalml.objectives.AUCWeighted` `class`
`(evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective`
`property), 315` `method), 402`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.MulticlassClassificationObjective` `(evalml.objectives.BalancedAccuracyBinary`
`property), 447` `class method), 404`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.objective_base.ObjectiveBase` `(evalml.objectives.BalancedAccuracyMulticlass`
`property), 318` `class method), 406`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.ObjectiveBase` `property),` `(evalml.objectives.binary_classification_objective.BinaryClassificatio`
`449` `class method), 305`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.regression_objective.RegressionObjective` `(evalml.objectives.BinaryClassificationObjective`
`property), 320` `class method), 408`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.RegressionObjective` `property),` `(evalml.objectives.cost_benefit_matrix.CostBenefitMatrix`
`467` `class method), 308`
`is_bounded_like_percentage()` `is_defined_for_problem_type()`
`(evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective`
`property), 388` `(evalml.objectives.CostBenefitMatrix` `class`
`is_cancelled()` (`evalml.automl.engine.cf_engine.CFComputation` `is_defined_for_problem_type()`
`property), 177` `(evalml.objectives.ExpVariance` `class method),`
`413`
`is_cancelled()` (`evalml.automl.engine.dask_engine.DaskComputation` `is_defined_for_problem_type()`
`property), 179` `(evalml.objectives.F1` `class method), 415`

<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.FIMacro</code> class method), 417	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MulticlassClassificationObjective</code> class method), 447
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.FIMicro</code> class method), 418	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.objective_base.ObjectiveBase</code> class method), 318
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.FIWeighted</code> class method), 420	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.ObjectiveBase</code> class method), 449
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.fraud_cost.FraudCost</code> class method), 310	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.Precision</code> class method), 452
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.FraudCost</code> class method), 422	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionMacro</code> class method), 454
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.Gini</code> class method), 425	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionMicro</code> class method), 455
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.lead_scoring.LeadScoring</code> class method), 313	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionWeighted</code> class method), 457
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LeadScoring</code> class method), 428	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.R2</code> class method), 458
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LogLossBinary</code> class method), 430	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.Recall</code> class method), 460
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LogLossMulticlass</code> class method), 432	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RecallMacro</code> class method), 462
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MAE</code> class method), 433	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RecallMicro</code> class method), 464
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MAPE</code> class method), 435	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RecallWeighted</code> class method), 465
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MaxError</code> class method), 437	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.regression_objective.RegressionObjective</code> class method), 320
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MCCBinary</code> class method), 439	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RegressionObjective</code> class method), 467
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MCCMulticlass</code> class method), 441	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RootMeanSquaredError</code> class method), 469
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MeanSquaredLogError</code> class method), 442	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RootMeanSquaredLogError</code> class method), 471
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MedianAE</code> class method), 444	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.sensitivity_low_alert.SensitivityLowAlert</code> class method), 323
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MSE</code> class method), 445	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.SensitivityLowAlert</code> class method), 323
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective</code> class method), 315	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MulticlassClassificationObjective</code> class method), 447

<i>(evalml.objectives.standard_metrics.AccuracyBinary class method), 326</i>	<i>(evalml.objectives.standard_metrics.MaxError class method), 358</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AccuracyMulticlass class method), 328</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MCCBinary class method), 360</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUC class method), 330</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MCCMulticlass class method), 362</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCMacro class method), 332</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MeanSquaredLogError class method), 364</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCMicro class method), 333</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MedianAE class method), 365</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCWeighted class method), 335</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MSE class method), 367</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.BalancedAccuracyBinary class method), 337</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.Precision class method), 369</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.BalancedAccuracyMulticlass class method), 339</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionMacro class method), 371</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.ExpVariance class method), 340</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionMicro class method), 372</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1 class method), 342</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionWeighted class method), 374</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1Macro class method), 344</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.R2 class method), 376</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1Micro class method), 346</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.Recall class method), 378</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1Weighted class method), 347</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.RecallMacro class method), 380</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.Gini class method), 349</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.RecallMicro class method), 381</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.LogLossBinary class method), 352</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.RecallWeighted class method), 383</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.LogLossMulticlass class method), 354</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.RootMeanSquaredError class method), 384</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MAE class method), 355</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.RootMeanSquaredLogError class method), 386</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MAPE class method), 357</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.time_series_regression_objective.TimeSeriesRegression class method), 388</i>
<i>is_defined_for_problem_type()</i>	<i>is_multiclass() (in module evalml.problem_types),</i>

1277			KNeighborsClassifier (class in evalml.pipelines.components), 980
is_multiclass() (in module evalml.problem_types.utils), 1275			KNeighborsClassifier (class in evalml.pipelines.components.estimators), 664
is_regression() (in module evalml.problem_types), 1277			KNeighborsClassifier (class in evalml.pipelines.components.estimators.classifiers), 538
is_regression() (in module evalml.problem_types.utils), 1275			KNeighborsClassifier (class in evalml.pipelines.components.estimators.classifiers.kneighbors_classifier), 506
is_search_space_exhausted() (evalml.tuners.grid_search_tuner.GridSearchTuner method), 1279			
is_search_space_exhausted() (evalml.tuners.GridSearchTuner method), 1284		L	
is_search_space_exhausted() (evalml.tuners.random_search_tuner.RandomSearchTuner method), 1280		LeadScoring (class in evalml.objectives), 426	
is_search_space_exhausted() (evalml.tuners.RandomSearchTuner method), 1285		LeadScoring (class in evalml.objectives.lead_scoring), 312	
is_search_space_exhausted() (evalml.tuners.skopt_tuner.SKOptTuner method), 1281		LightGBMClassifier (class in evalml.pipelines), 1174	
is_search_space_exhausted() (evalml.tuners.skopt_tuner.SKOptTuner method), 1281		LightGBMClassifier (class in evalml.pipelines.components), 983	
is_search_space_exhausted() (evalml.tuners.SKOptTuner method), 1286		LightGBMClassifier (class in evalml.pipelines.components.estimators), 667	
is_search_space_exhausted() (evalml.tuners.Tuner method), 1287		LightGBMClassifier (class in evalml.pipelines.components.estimators.classifiers), 541	
is_search_space_exhausted() (evalml.tuners.tuner.Tuner method), 1283		LightGBMClassifier (class in evalml.pipelines.components.estimators.classifiers.lightgbm_classifier), 510	
is_time_series() (in module evalml.problem_types), 1277		LightGBMRegressor (class in evalml.pipelines), 1177	
is_time_series() (in module evalml.problem_types.utils), 1275		LightGBMRegressor (class in evalml.pipelines.components), 986	
is_tree_estimator() (evalml.model_family.model_family.ModelFamily method), 269		LightGBMRegressor (class in evalml.pipelines.components.estimators), 670	
is_tree_estimator() (evalml.model_family.ModelFamily method), 271		LightGBMRegressor (class in evalml.pipelines.components.estimators.regressors), 610	
IterativeAlgorithm (class in evalml.automl.automl_algorithm), 174		LightGBMRegressor (class in evalml.pipelines.components.estimators.regressors.lightgbm_regressor), 574	
IterativeAlgorithm (class in evalml.automl.automl_algorithm.iterative_algorithm), 170		LinearDiscriminantAnalysis (class in evalml.pipelines.components), 989	
J		LinearDiscriminantAnalysis (class in evalml.pipelines.components.transformers), 871	
JobLogger (class in evalml.automl.engine.engine_base), 183		LinearDiscriminantAnalysis (class in evalml.pipelines.components.transformers.dimensionality_reducers), 703	
jupyter_check() (in module evalml.utils), 1298		LinearDiscriminantAnalysis (class in evalml.pipelines.components.transformers.dimensionality_reducers), 698	
jupyter_check() (in module evalml.utils.gen_utils), 1292		LinearRegressor (class in evalml.pipelines), 1180	
K			
KNeighborsClassifier (class in evalml.pipelines), 1172			

LinearRegressor	(class evalml.pipelines.components), 991	in load()	(evalml.pipelines.components.DropColumns static method), 956
LinearRegressor	(class evalml.pipelines.components.estimators), 673	in load()	(evalml.pipelines.components.DropNullColumns static method), 958
LinearRegressor	(class evalml.pipelines.components.estimators.regressors), 613	load()	(evalml.pipelines.components.DropRowsTransformer static method), 960
LinearRegressor	(class evalml.pipelines.components.estimators.regressors.linear_regressor), 577	load()	(evalml.pipelines.components.ElasticNetClassifier static method), 963
LinearRegressor	(class evalml.pipelines.components.estimators.regressors.linear_regressor), 577	load()	(evalml.pipelines.components.ElasticNetRegressor static method), 965
load()	(evalml.automl.automl_search.AutoMLSearch static method), 196	load()	(evalml.pipelines.components.EmailFeaturizer static method), 967
load()	(evalml.automl.AutoMLSearch static method), 207	load()	(evalml.pipelines.components.ensemble.sklearn_stacked_ensemble static method), 476
load()	(evalml.AutoMLSearch static method), 1303	load()	(evalml.pipelines.components.ensemble.sklearn_stacked_ensemble static method), 479
load()	(evalml.pipelines.ARIMARegressor static method), 1127	load()	(evalml.pipelines.components.ensemble.sklearn_stacked_ensemble static method), 482
load()	(evalml.pipelines.binary_classification_pipeline.BinaryClassifierPipeline static method), 1067	load()	(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier static method), 485
load()	(evalml.pipelines.BinaryClassificationPipeline static method), 1131	load()	(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier static method), 487
load()	(evalml.pipelines.CatBoostClassifier static method), 1135	load()	(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier static method), 490
load()	(evalml.pipelines.CatBoostRegressor static method), 1137	load()	(evalml.pipelines.components.Estimator static method), 969
load()	(evalml.pipelines.classification_pipeline.ClassificationPipeline static method), 1073	load()	(evalml.pipelines.components.estimators.ARIMARegressor static method), 634
load()	(evalml.pipelines.ClassificationPipeline static method), 1141	load()	(evalml.pipelines.components.estimators.BaselineClassifier static method), 636
load()	(evalml.pipelines.components.ARIMARegressor static method), 929	load()	(evalml.pipelines.components.estimators.BaselineRegressor static method), 638
load()	(evalml.pipelines.components.BaselineClassifier static method), 931	load()	(evalml.pipelines.components.estimators.CatBoostClassifier static method), 641
load()	(evalml.pipelines.components.BaselineRegressor static method), 933	load()	(evalml.pipelines.components.estimators.CatBoostRegressor static method), 644
load()	(evalml.pipelines.components.CatBoostClassifier static method), 936	load()	(evalml.pipelines.components.estimators.classifiers.baseline_class static method), 493
load()	(evalml.pipelines.components.CatBoostRegressor static method), 938	load()	(evalml.pipelines.components.estimators.classifiers.BaselineClass static method), 526
load()	(evalml.pipelines.components.component_base.ComponentBase static method), 919	load()	(evalml.pipelines.components.estimators.classifiers.catboost_class static method), 496
load()	(evalml.pipelines.components.ComponentBase static method), 940	load()	(evalml.pipelines.components.estimators.classifiers.CatBoostClass static method), 529
load()	(evalml.pipelines.components.DateTimeFeaturizer static method), 943	load()	(evalml.pipelines.components.estimators.classifiers.decision_tree_ static method), 499
load()	(evalml.pipelines.components.DecisionTreeClassifier static method), 946	load()	(evalml.pipelines.components.estimators.classifiers.DecisionTreeC static method), 532
load()	(evalml.pipelines.components.DecisionTreeRegressor static method), 949	load()	(evalml.pipelines.components.estimators.classifiers.elasticnet_cla static method), 502
load()	(evalml.pipelines.components.DelayedFeatureTransformer static method), 952	load()	(evalml.pipelines.components.estimators.classifiers.ElasticNetCla static method), 535
load()	(evalml.pipelines.components.DFSTransformer static method), 954	load()	(evalml.pipelines.components.estimators.classifiers.et_classifier.E static method), 505

load ()	(evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier static method), 538	(evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier static method), 682
load ()	(evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier static method), 508	(evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier static method), 684
load ()	(evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier static method), 541	(evalml.pipelines.components.estimators.regressors.arima_regressor static method), 558
load ()	(evalml.pipelines.components.estimators.classifiers.LightGBMClassifier static method), 512	(evalml.pipelines.components.estimators.regressors.ARIMARegressor static method), 596
load ()	(evalml.pipelines.components.estimators.classifiers.LightGBMClassifier static method), 544	(evalml.pipelines.components.estimators.regressors.baseline_regressor static method), 560
load ()	(evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier static method), 515	(evalml.pipelines.components.estimators.regressors.BaselineRegressor static method), 598
load ()	(evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier static method), 547	(evalml.pipelines.components.estimators.regressors.catboost_regressor static method), 563
load ()	(evalml.pipelines.components.estimators.classifiers.RandomForestClassifier static method), 549	(evalml.pipelines.components.estimators.regressors.CatBoostRegressor static method), 601
load ()	(evalml.pipelines.components.estimators.classifiers.rfc_classifier static method), 517	(evalml.pipelines.components.estimators.regressors.decision_tree_regressor static method), 567
load ()	(evalml.pipelines.components.estimators.classifiers.SVCClassifier static method), 520	(evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor static method), 604
load ()	(evalml.pipelines.components.estimators.classifiers.SVMClassifier static method), 552	(evalml.pipelines.components.estimators.regressors.elasticnet_regressor static method), 569
load ()	(evalml.pipelines.components.estimators.classifiers.xgboost_classifier static method), 523	(evalml.pipelines.components.estimators.regressors.ElasticNetRegressor static method), 606
load ()	(evalml.pipelines.components.estimators.classifiers.XGBoostClassifier static method), 554	(evalml.pipelines.components.estimators.regressors.et_regressor static method), 573
load ()	(evalml.pipelines.components.estimators.DecisionTreeClassifier static method), 647	(evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor static method), 609
load ()	(evalml.pipelines.components.estimators.DecisionTreeRegressor static method), 650	(evalml.pipelines.components.estimators.regressors.lightgbm_regressor static method), 576
load ()	(evalml.pipelines.components.estimators.ElasticNetClassifier static method), 653	(evalml.pipelines.components.estimators.regressors.LightGBMRegressor static method), 612
load ()	(evalml.pipelines.components.estimators.ElasticNetRegressor static method), 655	(evalml.pipelines.components.estimators.regressors.linear_regressor static method), 579
load ()	(evalml.pipelines.components.estimators.Estimator static method), 658	(evalml.pipelines.components.estimators.regressors.LinearRegressor static method), 615
load ()	(evalml.pipelines.components.estimators.estimator_factory static method), 630	(evalml.pipelines.components.estimators.regressors.prophet_regressor static method), 582
load ()	(evalml.pipelines.components.estimators.ExtraTreesClassifier static method), 660	(evalml.pipelines.components.estimators.regressors.ProphetRegressor static method), 617
load ()	(evalml.pipelines.components.estimators.ExtraTreesRegressor static method), 663	(evalml.pipelines.components.estimators.regressors.RandomForestRegressor static method), 620
load ()	(evalml.pipelines.components.estimators.KNeighborsClassifier static method), 666	(evalml.pipelines.components.estimators.regressors.rf_regressor static method), 584
load ()	(evalml.pipelines.components.estimators.LightGBMClassifier static method), 669	(evalml.pipelines.components.estimators.regressors.svm_regressor static method), 587
load ()	(evalml.pipelines.components.estimators.LightGBMRegressor static method), 672	(evalml.pipelines.components.estimators.regressors.SVMRegressor static method), 622
load ()	(evalml.pipelines.components.estimators.LinearRegressor static method), 674	(evalml.pipelines.components.estimators.regressors.time_series_baseline_regressor static method), 590
load ()	(evalml.pipelines.components.estimators.LogisticRegressionClassifier static method), 677	(evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineRegressor static method), 625
load ()	(evalml.pipelines.components.estimators.ProphetRegressor static method), 679	(evalml.pipelines.components.estimators.regressors.xgboost_regressor static method), 593

load()	(evalml.pipelines.components.estimators.regressors.XGBoostRegressor static method), 627	load()	(evalml.pipelines.components.RFRegressorSelectFromModel static method), 1024
load()	(evalml.pipelines.components.estimators.SVMClassifier static method), 686	load()	(evalml.pipelines.components.SelectByType static method), 1026
load()	(evalml.pipelines.components.estimators.SVMRegressor static method), 689	load()	(evalml.pipelines.components.SelectColumns static method), 1028
load()	(evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator static method), 691	load()	(evalml.pipelines.components.SimpleImputer static method), 1030
load()	(evalml.pipelines.components.estimators.XGBoostClassifier static method), 694	load()	(evalml.pipelines.components.SklearnStackedEnsembleClassifier static method), 1033
load()	(evalml.pipelines.components.estimators.XGBoostRegressor static method), 696	load()	(evalml.pipelines.components.SklearnStackedEnsembleRegressor static method), 1035
load()	(evalml.pipelines.components.ExtraTreesClassifier static method), 972	load()	(evalml.pipelines.components.StandardScaler static method), 1038
load()	(evalml.pipelines.components.ExtraTreesRegressor static method), 975	load()	(evalml.pipelines.components.SVMClassifier static method), 1040
load()	(evalml.pipelines.components.FeatureSelector static method), 977	load()	(evalml.pipelines.components.SVMRegressor static method), 1042
load()	(evalml.pipelines.components.Imputer static method), 980	load()	(evalml.pipelines.components.TargetEncoder static method), 1045
load()	(evalml.pipelines.components.KNeighborsClassifier static method), 982	load()	(evalml.pipelines.components.TargetImputer static method), 1047
load()	(evalml.pipelines.components.LightGBMClassifier static method), 985	load()	(evalml.pipelines.components.TextFeaturizer static method), 1049
load()	(evalml.pipelines.components.LightGBMRegressor static method), 988	load()	(evalml.pipelines.components.TimeSeriesBaselineEstimator static method), 1051
load()	(evalml.pipelines.components.LinearDiscriminantAnalysis static method), 990	load()	(evalml.pipelines.components.Transformer static method), 1054
load()	(evalml.pipelines.components.LinearRegressor static method), 992	load()	(evalml.pipelines.components.transformers.column_selectors.ColumnSelector static method), 838
load()	(evalml.pipelines.components.LogisticRegressionClassifier static method), 995	load()	(evalml.pipelines.components.transformers.column_selectors.DropColumns static method), 840
load()	(evalml.pipelines.components.LogTransformer static method), 997	load()	(evalml.pipelines.components.transformers.column_selectors.SelectColumns static method), 842
load()	(evalml.pipelines.components.LSA static method), 999	load()	(evalml.pipelines.components.transformers.column_selectors.SelectColumns static method), 844
load()	(evalml.pipelines.components.OneHotEncoder static method), 1002	load()	(evalml.pipelines.components.transformers.DateTimeFeaturizer static method), 853
load()	(evalml.pipelines.components.Oversampler static method), 1005	load()	(evalml.pipelines.components.transformers.DelayedFeatureTransformer static method), 855
load()	(evalml.pipelines.components.PCA static method), 1007	load()	(evalml.pipelines.components.transformers.DFSTransformer static method), 857
load()	(evalml.pipelines.components.PerColumnImputer static method), 1009	load()	(evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimension static method), 699
load()	(evalml.pipelines.components.PolynomialDetrender static method), 1011	load()	(evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimension static method), 704
load()	(evalml.pipelines.components.ProphetRegressor static method), 1014	load()	(evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimension static method), 706
load()	(evalml.pipelines.components.RandomForestClassifier static method), 1016	load()	(evalml.pipelines.components.transformers.dimensionality_reduction.ReduceDimension static method), 702
load()	(evalml.pipelines.components.RandomForestRegressor static method), 1018	load()	(evalml.pipelines.components.transformers.DropColumns static method), 860
load()	(evalml.pipelines.components.RFClassifierSelectFromModel static method), 1021	load()	(evalml.pipelines.components.transformers.DropNullColumns static method), 862

`load()` (`evalml.pipelines.components.transformers.DropRowsTransformer` static method), 864
`load()` (`evalml.pipelines.components.transformers.EmailFeaturizer` static method), 866
`load()` (`evalml.pipelines.components.transformers.encoder_one_hot` static method), 710
`load()` (`evalml.pipelines.components.transformers.encoder_one_hot_nullable` static method), 717
`load()` (`evalml.pipelines.components.transformers.encoder_target` static method), 714
`load()` (`evalml.pipelines.components.transformers.encoder_target_nullable` static method), 720
`load()` (`evalml.pipelines.components.transformers.feature_selection` static method), 722
`load()` (`evalml.pipelines.components.transformers.feature_selection_dataframe` static method), 732
`load()` (`evalml.pipelines.components.transformers.feature_selection_dataframe_nullable` static method), 725
`load()` (`evalml.pipelines.components.transformers.feature_selection_dataframe_nullable_nullable` static method), 729
`load()` (`evalml.pipelines.components.transformers.feature_selection_dataframe_nullable_nullable_nullable` static method), 734
`load()` (`evalml.pipelines.components.transformers.feature_selection_dataframe_nullable_nullable_nullable_nullable` static method), 737
`load()` (`evalml.pipelines.components.transformers.FeatureSelector` static method), 868
`load()` (`evalml.pipelines.components.transformers.Imputer` static method), 871
`load()` (`evalml.pipelines.components.transformers.imputer` static method), 752
`load()` (`evalml.pipelines.components.transformers.imputer_nullable` static method), 740
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable` static method), 743
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable_nullable` static method), 754
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable_nullable_nullable` static method), 745
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable_nullable_nullable_nullable` static method), 757
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable_nullable_nullable_nullable_nullable` static method), 748
`load()` (`evalml.pipelines.components.transformers.imputer_nullable_nullable_nullable_nullable_nullable_nullable_nullable` static method), 759
`load()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis` static method), 873
`load()` (`evalml.pipelines.components.transformers.LogTransformer` static method), 875
`load()` (`evalml.pipelines.components.transformers.LSA` static method), 877
`load()` (`evalml.pipelines.components.transformers.OneHotEncoder` static method), 880
`load()` (`evalml.pipelines.components.transformers.Oversampler` static method), 882
`load()` (`evalml.pipelines.components.transformers.PCA` static method), 885
`load()` (`evalml.pipelines.components.transformers.PerColumnImputer` static method), 887
`load()` (`evalml.pipelines.components.transformers.polynomial_detrender` static method), 889
`load()` (`evalml.pipelines.components.transformers.preprocessing.datetime` static method), 762
`load()` (`evalml.pipelines.components.transformers.preprocessing.DateTransformer` static method), 793
`load()` (`evalml.pipelines.components.transformers.preprocessing.delayed` static method), 765
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureSelector` static method), 795
`load()` (`evalml.pipelines.components.transformers.preprocessing.DFSTransformer` static method), 797
`load()` (`evalml.pipelines.components.transformers.preprocessing.RFClassifier` static method), 767
`load()` (`evalml.pipelines.components.transformers.preprocessing.RFRegressor` static method), 770
`load()` (`evalml.pipelines.components.transformers.preprocessing.DropNulls` static method), 800
`load()` (`evalml.pipelines.components.transformers.preprocessing.DropRows` static method), 802
`load()` (`evalml.pipelines.components.transformers.preprocessing.EmailFeaturizer` static method), 804
`load()` (`evalml.pipelines.components.transformers.preprocessing.feature` static method), 772
`load()` (`evalml.pipelines.components.transformers.preprocessing.log_transform` static method), 775
`load()` (`evalml.pipelines.components.transformers.preprocessing.LogTransformer` static method), 806
`load()` (`evalml.pipelines.components.transformers.preprocessing.LSA` static method), 808
`load()` (`evalml.pipelines.components.transformers.preprocessing.lsa.LSA` static method), 777
`load()` (`evalml.pipelines.components.transformers.preprocessing.polynomial` static method), 780
`load()` (`evalml.pipelines.components.transformers.preprocessing.Polynomial` static method), 810
`load()` (`evalml.pipelines.components.transformers.preprocessing.text_featurizer` static method), 783
`load()` (`evalml.pipelines.components.transformers.preprocessing.text_transformer` static method), 785
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextFeaturizer` static method), 813
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextTransformer` static method), 815
`load()` (`evalml.pipelines.components.transformers.preprocessing.transformer` static method), 787
`load()` (`evalml.pipelines.components.transformers.preprocessing.transformer_nullable` static method), 789
`load()` (`evalml.pipelines.components.transformers.preprocessing.URLFeaturizer` static method), 817

`load()` (`evalml.pipelines.components.transformers.RFClassifierSelectFromModel` static method), 892
`load()` (`evalml.pipelines.components.transformers.RFRegressorSelectFromModel` static method), 895
`load()` (`evalml.pipelines.components.transformers.sampler.base.SamplerBase` static method), 819
`load()` (`evalml.pipelines.components.transformers.sampler.OverSampler` static method), 828
`load()` (`evalml.pipelines.components.transformers.sampler.OverSampler` static method), 822
`load()` (`evalml.pipelines.components.transformers.sampler.UnderSampler` static method), 830
`load()` (`evalml.pipelines.components.transformers.sampler.UnderSampler` static method), 825
`load()` (`evalml.pipelines.components.transformers.scalers.standard_scaler.StandardScaler` static method), 833
`load()` (`evalml.pipelines.components.transformers.scalers.standard_scaler.StandardScaler` static method), 835
`load()` (`evalml.pipelines.components.transformers.SelectByType` static method), 897
`load()` (`evalml.pipelines.components.transformers.SelectColumns` static method), 899
`load()` (`evalml.pipelines.components.transformers.SimpleImputer` static method), 901
`load()` (`evalml.pipelines.components.transformers.StandardScaler` static method), 903
`load()` (`evalml.pipelines.components.transformers.TargetEncoder` static method), 906
`load()` (`evalml.pipelines.components.transformers.TargetImputer` static method), 908
`load()` (`evalml.pipelines.components.transformers.TextFeaturizer` static method), 910
`load()` (`evalml.pipelines.components.transformers.Transformer` static method), 912
`load()` (`evalml.pipelines.components.transformers.transformer.TargetTransformer` static method), 846
`load()` (`evalml.pipelines.components.transformers.transformer.TargetTransformer` static method), 849
`load()` (`evalml.pipelines.components.transformers.Undersampler` static method), 915
`load()` (`evalml.pipelines.components.transformers.URLFeaturizer` static method), 917
`load()` (`evalml.pipelines.components.Undersampler` static method), 1056
`load()` (`evalml.pipelines.components.URLFeaturizer` static method), 1058
`load()` (`evalml.pipelines.components.XGBoostClassifier` static method), 1060
`load()` (`evalml.pipelines.components.XGBoostRegressor` static method), 1063
`load()` (`evalml.pipelines.DecisionTreeClassifier` static method), 1148
`load()` (`evalml.pipelines.DecisionTreeRegressor` static method), 1151
`load()` (`evalml.pipelines.DelayedFeatureTransformer` static method), 1153
`load()` (`evalml.pipelines.components.transformers.DFSTransformer` static method), 1156
`load()` (`evalml.pipelines.components.transformers.ElasticNetClassifier` static method), 1158
`load()` (`evalml.pipelines.components.transformers.ElasticNetRegressor` static method), 1160
`load()` (`evalml.pipelines.components.transformers.Estimator` static method), 1163
`load()` (`evalml.pipelines.components.transformers.ExtraTreesClassifier` static method), 1166
`load()` (`evalml.pipelines.components.transformers.ExtraTreesRegressor` static method), 1169
`load()` (`evalml.pipelines.components.transformers.FeatureSelector` static method), 1171
`load()` (`evalml.pipelines.components.transformers.KNeighborsClassifier` static method), 1174
`load()` (`evalml.pipelines.components.transformers.LightGBMClassifier` static method), 1177
`load()` (`evalml.pipelines.components.transformers.LightGBMRegressor` static method), 1179
`load()` (`evalml.pipelines.components.transformers.LinearRegressor` static method), 1181
`load()` (`evalml.pipelines.components.transformers.LogisticRegressionClassifier` static method), 1184
`load()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassifier` static method), 1082
`load()` (`evalml.pipelines.MulticlassClassificationPipeline` static method), 1188
`load()` (`evalml.pipelines.OneHotEncoder` static method), 1192
`load()` (`evalml.pipelines.PerColumnImputer` static method), 1194
`load()` (`evalml.pipelines.pipeline_base.PipelineBase` static method), 1087
`load()` (`evalml.pipelines.PipelineBase` static method), 1198
`load()` (`evalml.pipelines.ProphetRegressor` static method), 1201
`load()` (`evalml.pipelines.RandomForestClassifier` static method), 1203
`load()` (`evalml.pipelines.RandomForestRegressor` static method), 1205
`load()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` static method), 1092
`load()` (`evalml.pipelines.RegressionPipeline` static method), 1209
`load()` (`evalml.pipelines.RFClassifierSelectFromModel` static method), 1212
`load()` (`evalml.pipelines.RFRegressorSelectFromModel` static method), 1215
`load()` (`evalml.pipelines.SimpleImputer` static method), 1217

[load\(\)](#) ([evalml.pipelines.SklearnStackedEnsembleClassifier](#) static method), 1220
[load\(\)](#) ([evalml.pipelines.SklearnStackedEnsembleRegressor](#) static method), 1222
[load\(\)](#) ([evalml.pipelines.StandardScaler](#) static method), 1224
[load\(\)](#) ([evalml.pipelines.SVMClassifier](#) static method), 1227
[load\(\)](#) ([evalml.pipelines.SVMRegressor](#) static method), 1229
[load\(\)](#) ([evalml.pipelines.TargetEncoder](#) static method), 1232
[load\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) static method), 1098
[load\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline](#) static method), 1104
[load\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline](#) static method), 1109
[load\(\)](#) ([evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase](#) static method), 1115
[load\(\)](#) ([evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline](#) static method), 1120
[load\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) static method), 1235
[load\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#) static method), 1241
[load\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#) static method), 1247
[load\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#) static method), 1252
[load\(\)](#) ([evalml.pipelines.Transformer](#) static method), 1256
[load\(\)](#) ([evalml.pipelines.XGBoostClassifier](#) static method), 1258
[load\(\)](#) ([evalml.pipelines.XGBoostRegressor](#) static method), 1260
[load_breast_cancer\(\)](#) (in module [evalml.demos](#)), 263
[load_breast_cancer\(\)](#) (in module [evalml.demos.breast_cancer](#)), 261
[load_churn\(\)](#) (in module [evalml.demos](#)), 263
[load_churn\(\)](#) (in module [evalml.demos.churn](#)), 261
[load_data\(\)](#) (in module [evalml.preprocessing](#)), 1270
[load_data\(\)](#) (in module [evalml.preprocessing.utils](#)), 1269
[load_diabetes\(\)](#) (in module [evalml.demos](#)), 264
[load_diabetes\(\)](#) (in module [evalml.demos.diabetes](#)), 262
[load_fraud\(\)](#) (in module [evalml.demos](#)), 264
[load_fraud\(\)](#) (in module [evalml.demos.fraud](#)), 262
[load_weather\(\)](#) (in module [evalml.demos](#)), 264
[load_weather\(\)](#) (in module [evalml.demos.weather](#)), 263
[load_wine\(\)](#) (in module [evalml.demos](#)), 264
[load_wine\(\)](#) (in module [evalml.demos.wine](#)), 263
[log_error_callback\(\)](#) (in module [evalml.callbacks](#)), 199
[log_subtitle\(\)](#) (in module [evalml.utils](#)), 1298
[log_subtitle\(\)](#) (in module [evalml.utils.logger](#)), 1293
[log_title\(\)](#) (in module [evalml.utils](#)), 1298
[log_title\(\)](#) (in module [evalml.utils.logger](#)), 1293
[logger](#) (in module [evalml.callbacks](#)), 199
[logger](#) (in module [evalml.objectives.sensitivity_low_alert](#)), 322
[logger](#) (in module [evalml.pipelines.pipeline_base](#)), 1124
[logger](#) (in module [evalml.pipelines.utils](#)), 1123
[logger](#) (in module [evalml.utils.gen_utils](#)), 1292
[LogLossBinary](#) (class in [evalml.objectives](#)), 429
[LogLossBinary](#) (class in [evalml.objectives.standard_metrics](#)), 350
[LogLossMulticlass](#) (class in [evalml.objectives](#)), 431
[LogLossMulticlass](#) (class in [evalml.objectives.standard_metrics](#)), 353
[LogTransformer](#) (class in [evalml.pipelines.components](#)), 996
[LogTransformer](#) (class in [evalml.pipelines.components.transformers](#)), 873
[LogTransformer](#) (class in [evalml.pipelines.components.transformers.preprocessing](#)), 804
[LogTransformer](#) (class in [evalml.pipelines.components.transformers.preprocessing.log_transformer](#)), 773
[LSA](#) (class in [evalml.pipelines.components](#)), 998
[LSA](#) (class in [evalml.pipelines.components.transformers](#)), 876
[LSA](#) (class in [evalml.pipelines.components.transformers.preprocessing](#)), 807
[LSA](#) (class in [evalml.pipelines.components.transformers.preprocessing.lsa](#)), 776

M

MAE (class in `evalml.objectives`), 432

MAE (class in `evalml.objectives.standard_metrics`), 354

`make_balancing_dictionary()` (in module `evalml.pipelines.components.utils`), 922

`make_data_splitter()` (in module `evalml.automl`), 209

`make_data_splitter()` (in module `evalml.automl.utils`), 202

`make_pipeline()` (in module `evalml.pipelines.utils`), 1123

`make_timeseries_baseline_pipeline()` (in module `evalml.pipelines.utils`), 1123

MAPE (class in `evalml.objectives`), 434

MAPE (class in `evalml.objectives.standard_metrics`), 356

MaxError (class in `evalml.objectives`), 436

MaxError (class in `evalml.objectives.standard_metrics`), 358

MCCBinary (class in `evalml.objectives`), 437

MCCBinary (class in `evalml.objectives.standard_metrics`), 359

MCCMulticlass (class in `evalml.objectives`), 440

MCCMulticlass (class in `evalml.objectives.standard_metrics`), 361

MeanSquaredLogError (class in `evalml.objectives`), 441

MeanSquaredLogError (class in `evalml.objectives.standard_metrics`), 363

MedianAE (class in `evalml.objectives`), 443

MedianAE (class in `evalml.objectives.standard_metrics`), 365

`method` (in module `evalml.utils.update_checker`), 1294

MethodPropertyNotFoundError, 265, 267

MissingComponentError, 265, 267

`model_family()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` property), 1067

`model_family()` (`evalml.pipelines.BinaryClassificationPipeline` property), 1131

`model_family()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` property), 1073

`model_family()` (`evalml.pipelines.ClassificationPipeline` property), 1141

`model_family()` (`evalml.pipelines.components.component_base.ComponentBase` property), 919

`model_family()` (`evalml.pipelines.components.ComponentBase` property), 940

`model_family()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` property), 1082

`model_family()` (`evalml.pipelines.MulticlassClassificationPipeline` property), 1188

`model_family()` (`evalml.pipelines.pipeline_base.PipelineBase` property), 1087

`model_family()` (`evalml.pipelines.PipelineBase` property), 1198

`model_family()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` property), 1093

`model_family()` (`evalml.pipelines.RegressionPipeline` property), 1209

`model_family()` (`evalml.pipelines.time_series_classification_pipelines` property), 1098

`model_family()` (`evalml.pipelines.time_series_classification_pipelines` property), 1104

`model_family()` (`evalml.pipelines.time_series_classification_pipelines` property), 1109

`model_family()` (`evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase` property), 1115

`model_family()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` property), 1120

`model_family()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` property), 1235

`model_family()` (`evalml.pipelines.TimeSeriesClassificationPipeline` property), 1241

`model_family()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` property), 1247

`model_family()` (`evalml.pipelines.TimeSeriesRegressionPipeline` property), 1252

ModelFamily (class in `evalml.model_family`), 270

ModelFamily (class in `evalml.model_family.model_family`), 269

`modifies_features()` (`evalml.pipelines.components.component_base.ComponentBase` property), 919

`modifies_features()` (`evalml.pipelines.components.ComponentBase` property), 940

`modifies_target()` (`evalml.pipelines.components.component_base.ComponentBase` property), 919

`modifies_target()` (`evalml.pipelines.components.ComponentBase` property), 941

module

`evalml`, 166

`evalml.automl`, 166

`evalml.automl.automl_algorithm`, 166

`evalml.automl.automl_algorithm.automl_algorithm`, 166

`evalml.automl.automl_algorithm.default_algorithm`, 167

`evalml.automl.automl_algorithm.iterative_algorithm`, 169

`evalml.automl.automl_search`, 191

`evalml.automl.callbacks`, 199

`evalml.automl.engine`, 176

`evalml.automl.engine.cf_engine`, 176

`evalml.automl.engine.dask_engine`, 179

`evalml.automl.engine.engine_base`, 179

[181](#)
[evalml.automl.engine.sequential_engine,](#)
[184](#)
[evalml.automl.pipeline_search_plots,](#)
[200](#)
[evalml.automl.utils,](#) [200](#)
[evalml.data_checks,](#) [212](#)
[evalml.data_checks.class_imbalance_data_check,](#)
[212](#)
[evalml.data_checks.data_check,](#) [213](#)
[evalml.data_checks.data_check_action,](#)
[214](#)
[evalml.data_checks.data_check_action_code,](#)
[215](#)
[evalml.data_checks.data_check_message,](#)
[216](#)
[evalml.data_checks.data_check_message_code,](#)
[217](#)
[evalml.data_checks.data_check_message_type,](#)
[219](#)
[evalml.data_checks.data_checks,](#) [220](#)
[evalml.data_checks.datetime_format_data_check,](#)
[220](#)
[evalml.data_checks.datetime_nan_data_check,](#)
[222](#)
[evalml.data_checks.default_data_checks,](#)
[223](#)
[evalml.data_checks.highly_null_data_check,](#)
[224](#)
[evalml.data_checks.id_columns_data_check,](#)
[226](#)
[evalml.data_checks.invalid_targets_data_check,](#)
[227](#)
[evalml.data_checks.multicollinearity_data_check,](#)
[228](#)
[evalml.data_checks.natural_language_nan_data_check,](#)
[229](#)
[evalml.data_checks.no_variance_data_check,](#)
[231](#)
[evalml.data_checks.outliers_data_check,](#)
[232](#)
[evalml.data_checks.sparsity_data_check,](#)
[233](#)
[evalml.data_checks.target_distribution_data_check,](#)
[235](#)
[evalml.data_checks.target_leakage_data_check,](#)
[236](#)
[evalml.data_checks.uniqueness_data_check,](#)
[237](#)
[evalml.data_checks.utils,](#) [239](#)
[evalml.demos,](#) [260](#)
[evalml.demos.breast_cancer,](#) [260](#)
[evalml.demos.churn,](#) [261](#)
[evalml.demos.diabetes,](#) [261](#)
[evalml.demos.fraud,](#) [262](#)
[evalml.demos.weather,](#) [262](#)
[evalml.demos.wine,](#) [263](#)
[evalml.exceptions,](#) [264](#)
[evalml.exceptions.exceptions,](#) [264](#)
[evalml.model_family,](#) [269](#)
[evalml.model_family.model_family,](#)
[evalml.model_family.utils,](#) [270](#)
[evalml.model_understanding,](#) [271](#)
[evalml.model_understanding.force_plots,](#)
[278](#)
[evalml.model_understanding.graphs,](#)
[279](#)
[evalml.model_understanding.permutation_importance,](#)
[289](#)
[evalml.model_understanding.prediction_explanations,](#)
[271](#)
[evalml.model_understanding.prediction_explanations,](#)
[271](#)
[evalml.objectives,](#) [303](#)
[evalml.objectives.binary_classification_objective,](#)
[303](#)
[evalml.objectives.cost_benefit_matrix,](#)
[306](#)
[evalml.objectives.fraud_cost,](#) [309](#)
[evalml.objectives.lead_scoring,](#) [312](#)
[evalml.objectives.multiclass_classification_objective,](#)
[314](#)
[evalml.objectives.objective_base,](#)
[317](#)
[evalml.objectives.regression_objective,](#)
[319](#)
[evalml.objectives.sensitivity_low_alert,](#)
[321](#)
[evalml.objectives.standard_metrics,](#)
[324](#)
[evalml.objectives.time_series_regression_objective,](#)
[387](#)
[evalml.objectives.utils,](#) [389](#)
[evalml.pipelines,](#) [474](#)
[evalml.pipelines.binary_classification_pipeline,](#)
[1064](#)
[evalml.pipelines.binary_classification_pipeline,](#)
[1069](#)
[evalml.pipelines.classification_pipeline,](#)
[1070](#)
[evalml.pipelines.component_graph,](#)
[1075](#)
[evalml.pipelines.components,](#) [474](#)
[evalml.pipelines.components.component_base,](#)
[918](#)
[evalml.pipelines.components.component_base_meta,](#)
[920](#)

evalml.pipelines.components.ensemble,	evalml.pipelines.components.estimators.regresso
474	580
evalml.pipelines.components.ensemble.sklearnstackedensemble,	evalml.pipelines.components.estimators.regresso
474	583
evalml.pipelines.components.ensemble.sklearnstackedensemble,	evalml.pipelines.components.estimators.regresso
477	585
evalml.pipelines.components.ensemble.sklearnstackedensemble,	evalml.pipelines.components.estimators.regresso
480	588
evalml.pipelines.components.estimators,	evalml.pipelines.components.estimators.regresso
491	591
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers,
491	697
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.column
491	836
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.dimens
494	697
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.dimens
497	697
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.dimens
500	700
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.encode
503	707
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.encode
506	707
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.encode
509	711
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.feature
513	720
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.feature
516	720
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.feature
518	723
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.feature
521	726
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.impute
628	738
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.impute
555	738
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.impute
555	741
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.impute
559	744
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.impute
561	746
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.prepro
564	760
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.prepro
568	760
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.prepro
570	763
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.prepro
574	766
evalml.pipelines.components.estimators.categorical,	evalml.pipelines.components.transformers.prepro
577	768

[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays](#), 1269
[771](#) [evalml.problem_types](#), 1273
[evalml.pipelines.components.transformer_separators.problem_types.problem_types](#),
[773](#) [1273](#)
[evalml.pipelines.components.transformer_separators.problem_types.utils](#), 1274
[776](#) [evalml.tuners](#), 1278
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[778](#) [1278](#)
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[781](#) [1279](#)
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[783](#) [evalml.tuners.tuner](#), 1282
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[786](#) [evalml.utils](#), 1288
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[818](#) [evalml.utils.base_meta](#), 1288
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[818](#) [evalml.utils.cli_utils](#), 1288
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[818](#) [evalml.utils.logger](#), 1293
[evalml.pipelines.components.transformer_separators.preprocessing_features_to_arrays_tuner](#),
[820](#) [evalml.utils.woodwork_utils](#), 1294
[evalml.pipelines.components.transformers](#), 1294
[823](#) [MSE \(class in evalml.objectives\)](#), 445,
[evalml.pipelines.components.transformers](#), 366
[831](#) [MulticlassClassificationObjective \(class](#)
[in evalml.objectives\)](#), 446
[evalml.pipelines.components.transformers](#), 314
[831](#) [MulticlassClassificationObjective \(class](#)
[in evalml.objectives.multiclass_classification_objective\)](#),
[314](#)
[evalml.pipelines.components.transformers](#), 1185
[845](#) [MulticlassClassificationPipeline \(class in](#)
[evalml.pipelines\)](#), 1185
[evalml.pipelines.components.utils](#), 1079
[921](#) [MulticlassClassificationPipeline \(class in](#)
[evalml.pipelines.multiclass_classification_pipeline\)](#),
[1079](#)
[evalml.pipelines.pipeline_base](#), 1084 [MulticollinearityDataCheck \(class in](#)
[evalml.pipelines.pipeline_meta](#), 1089 [evalml.data_checks\)](#), 253
[evalml.pipelines.regression_pipeline](#), 1089 [MulticollinearityDataCheck \(class in](#)
[evalml.pipelines.time_series_classification_pipeline](#), 1094
[1094](#) [evalml.pipelines.time_series_classification_pipeline](#),
[1112](#) [name \(\) \(evalml.data_checks.class_imbalance_data_check.ClassImbalanceDataCheck](#)
[method\)](#), 212
[evalml.pipelines.time_series_regression_pipeline](#), 1117
[1117](#) [name \(\) \(evalml.data_checks.ClassImbalanceDataCheck](#)
[method\)](#), 241
[evalml.pipelines.utils](#), 1122
[evalml.preprocessing](#), 1261
[evalml.preprocessing.data_splitters](#), 1261
[1261](#) [name \(\) \(evalml.data_checks.data_check.DataCheck](#)
[method\)](#), 214
[evalml.preprocessing.data_splitters.balanced_classification_sampler](#), 1261
[1261](#) [name \(\) \(evalml.data_checks.data_check_message_code.DataCheckMessageCode](#)
[method\)](#), 219
[evalml.preprocessing.data_splitters.sampler](#), 1263
[1263](#) [name \(\) \(evalml.data_checks.data_check_message_type.DataCheckMessageType](#)
[method\)](#), 219
[evalml.preprocessing.data_splitters.time_series_sampler](#), 1263
[1263](#) [name \(\) \(evalml.data_checks.DataCheck method\)](#), 242
[evalml.preprocessing.data_splitters.training_validation_splitter](#), 1265
[1265](#) [name \(\) \(evalml.data_checks.DataCheckActionCode](#)
[method\)](#), 243

<code>name()</code> (<code>evalml.data_checks.DataCheckMessageCode</code> method), 245	<code>name()</code> (<code>evalml.data_checks.UniquenessDataCheck</code> method), 259
<code>name()</code> (<code>evalml.data_checks.DataCheckMessageType</code> method), 246	<code>name()</code> (<code>evalml.exceptions.exceptions.PartialDependenceErrorCode</code> method), 266
<code>name()</code> (<code>evalml.data_checks.datetime_format_data_check.DateTimeFormatDataCheck</code> method), 221	<code>name()</code> (<code>evalml.exceptions.exceptions.PartialDependenceErrorCode</code> method), 268
<code>name()</code> (<code>evalml.data_checks.datetime_nan_data_check.DateTimeNaNDataCheck</code> method), 222	<code>name()</code> (<code>evalml.model_family.model_family.ModelFamily</code> method), 270
<code>name()</code> (<code>evalml.data_checks.DateTimeFormatDataCheck</code> method), 247	<code>name()</code> (<code>evalml.model_family.ModelFamily</code> method), 271
<code>name()</code> (<code>evalml.data_checks.DateTimeNaNDataCheck</code> method), 247	<code>name()</code> (<code>evalml.model_understanding.prediction_explanations.explainers</code> method), 275
<code>name()</code> (<code>evalml.data_checks.highly_null_data_check.HighlyNullDataCheck</code> method), 225	<code>name()</code> (<code>evalml.objectives.binary_classification_objective.BinaryClassificationObjective</code> property), 305
<code>name()</code> (<code>evalml.data_checks.HighlyNullDataCheck</code> method), 250	<code>name()</code> (<code>evalml.objectives.BinaryClassificationObjective</code> property), 408
<code>name()</code> (<code>evalml.data_checks.id_columns_data_check.IDColumnsDataCheck</code> method), 226	<code>name()</code> (<code>evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective</code> property), 316
<code>name()</code> (<code>evalml.data_checks.IDColumnsDataCheck</code> method), 251	<code>name()</code> (<code>evalml.objectives.MulticlassClassificationObjective</code> property), 447
<code>name()</code> (<code>evalml.data_checks.invalid_targets_data_check.InvalidTargetsDataCheck</code> method), 228	<code>name()</code> (<code>evalml.objectives.objective_base.ObjectiveBase</code> property), 318
<code>name()</code> (<code>evalml.data_checks.InvalidTargetDataCheck</code> method), 252	<code>name()</code> (<code>evalml.objectives.ObjectiveBase</code> property), 450
<code>name()</code> (<code>evalml.data_checks.multicollinearity_data_check.MulticollinearityDataCheck</code> method), 229	<code>name()</code> (<code>evalml.objectives.regression_objective.RegressionObjective</code> property), 467
<code>name()</code> (<code>evalml.data_checks.MulticollinearityDataCheck</code> method), 253	<code>name()</code> (<code>evalml.objectives.RegressionObjective</code> property), 467
<code>name()</code> (<code>evalml.data_checks.natural_language_nan_data_check.NaturalLanguageNaNDataCheck</code> method), 230	<code>name()</code> (<code>evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective</code> property), 488
<code>name()</code> (<code>evalml.data_checks.NaturalLanguageNaNDataCheck</code> method), 254	<code>name()</code> (<code>evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline</code> property), 1067
<code>name()</code> (<code>evalml.data_checks.no_variance_data_check.NoVarianceDataCheck</code> method), 231	<code>name()</code> (<code>evalml.pipelines.BinaryClassificationPipeline</code> property), 1131
<code>name()</code> (<code>evalml.data_checks.NoVarianceDataCheck</code> method), 254	<code>name()</code> (<code>evalml.pipelines.classification_pipeline.ClassificationPipeline</code> property), 1073
<code>name()</code> (<code>evalml.data_checks.outliers_data_check.OutliersDataCheck</code> method), 232	<code>name()</code> (<code>evalml.pipelines.ClassificationPipeline</code> property), 1141
<code>name()</code> (<code>evalml.data_checks.OutliersDataCheck</code> method), 255	<code>name()</code> (<code>evalml.pipelines.components.component_base.ComponentBase</code> property), 919
<code>name()</code> (<code>evalml.data_checks.sparsity_data_check.SparsityDataCheck</code> method), 233	<code>name()</code> (<code>evalml.pipelines.components.ComponentBase</code> property), 941
<code>name()</code> (<code>evalml.data_checks.SparsityDataCheck</code> method), 256	<code>name()</code> (<code>evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.SklearnStackedEnsembleClassifier</code> property), 476
<code>name()</code> (<code>evalml.data_checks.target_distribution_data_check.TargetDistributionDataCheck</code> method), 235	<code>name()</code> (<code>evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier</code> property), 476
<code>name()</code> (<code>evalml.data_checks.target_leakage_data_check.TargetLeakageDataCheck</code> method), 236	<code>name()</code> (<code>evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier</code> property), 476
<code>name()</code> (<code>evalml.data_checks.TargetDistributionDataCheck</code> method), 257	<code>name()</code> (<code>evalml.pipelines.components.estimator.Estimator</code> property), 658
<code>name()</code> (<code>evalml.data_checks.TargetLeakageDataCheck</code> method), 258	<code>name()</code> (<code>evalml.pipelines.components.estimator.estimator.Estimator</code> property), 630
<code>name()</code> (<code>evalml.data_checks.uniqueness_data_check.UniquenessDataCheck</code> method), 238	<code>name()</code> (<code>evalml.pipelines.components.FeatureSelector</code> property), 977
	<code>name()</code> (<code>evalml.pipelines.components.Transformer</code> property), 977

property), 1054

name () (evalml.pipelines.components.transformers.column_selector.ColumnSelector property), 838

name () (evalml.pipelines.components.transformers.feature_selection.FeatureSelector property), 722

name () (evalml.pipelines.components.transformers.feature_selection.FeatureSelector evalml.data_checks.natural_language_nan_data_check), 868

name () (evalml.pipelines.components.transformers.preprocessing.TextTransformer evalml.pipelines.CatBoostClassifier method), 1127

name () (evalml.pipelines.components.transformers.preprocessing.TextTransformer evalml.pipelines.CatBoostRegressor method), 1135

name () (evalml.pipelines.components.transformers.sampler_base.SamplerBase evalml.pipelines.CatBoostRegressor method), 1137

name () (evalml.pipelines.components.transformers.Transformers_fitting () (evalml.pipelines.components.ARIMARegressor method), 929

name () (evalml.pipelines.components.transformers.transformer_base.Transformers_fitting () (evalml.pipelines.components.BaselineClassifier method), 931

name () (evalml.pipelines.components.transformers.transformer_base.Transformers_fitting () (evalml.pipelines.components.BaselineRegressor method), 934

name () (evalml.pipelines.Estimator property), 1163

name () (evalml.pipelines.FeatureSelector property), 1171

name () (evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline property), 1082

name () (evalml.pipelines.MulticlassClassificationPipeline property), 1188

name () (evalml.pipelines.pipeline_base.PipelineBase property), 1087

name () (evalml.pipelines.PipelineBase property), 1198

name () (evalml.pipelines.regression_pipeline.RegressionPipeline property), 1093

name () (evalml.pipelines.RegressionPipeline property), 1209

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesBinaryClassificationPipeline property), 1098

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1104

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline property), 1109

name () (evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase property), 1115

name () (evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline property), 1120

name () (evalml.pipelines.TimeSeriesBinaryClassificationPipeline property), 1235

name () (evalml.pipelines.TimeSeriesClassificationPipeline property), 1241

name () (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property), 1247

name () (evalml.pipelines.TimeSeriesRegressionPipeline property), 1252

name () (evalml.pipelines.Transformer property), 1256

name () (evalml.problem_types.problem_types.ProblemTypes method), 273

name () (evalml.problem_types.ProblemTypes method), 273

NaturalLanguageNaNDataCheck (class in evalml.data_checks.natural_language_nan_data_check), 230

evalml.pipelines.ARIMARegressor method), 1127

evalml.pipelines.CatBoostClassifier method), 1135

evalml.pipelines.CatBoostRegressor method), 1137

evalml.pipelines.components.ARIMARegressor method), 929

evalml.pipelines.components.BaselineClassifier method), 931

evalml.pipelines.components.BaselineRegressor method), 934

needs_fitting () (evalml.pipelines.components.CatBoostClassifier method), 936

needs_fitting () (evalml.pipelines.components.CatBoostRegressor method), 936

needs_fitting () (evalml.pipelines.components.component_base.ComponentBase method), 919

needs_fitting () (evalml.pipelines.components.ComponentBase method), 941

needs_fitting () (evalml.pipelines.components.DateTimeFeaturizer method), 943

needs_fitting () (evalml.pipelines.components.DecisionTreeClassifier method), 946

needs_fitting () (evalml.pipelines.components.DecisionTreeRegressor method), 949

evalml.pipelines.components.DFSTransformer method), 954

evalml.pipelines.components.DropNullColumns method), 958

evalml.pipelines.components.DropRowsTransformer method), 960

evalml.pipelines.components.ElasticNetClassifier method), 963

evalml.pipelines.components.ElasticNetRegressor method), 965

evalml.pipelines.components.EmailFeaturizer method), 967

evalml.pipelines.components.ensemble.sklearn_stack method), 476

evalml.pipelines.components.ensemble.sklearn_stack method), 479

evalml.pipelines.components.ensemble.sklearn_stack method), 482

evalml.pipelines.components.ensemble.SklearnStack method), 482

[illegible]

`method`), 1220
`needs_fitting()` (`evalml.pipelines.SklearnStackedEnsembleRegressor` `method`), 167
`method`), 1222
`needs_fitting()` (`evalml.pipelines.StandardScaler` `method`), 1224
`needs_fitting()` (`evalml.pipelines.SVMClassifier` `method`), 1227
`needs_fitting()` (`evalml.pipelines.SVMRegressor` `method`), 1229
`needs_fitting()` (`evalml.pipelines.TargetEncoder` `method`), 1232
`needs_fitting()` (`evalml.pipelines.Transformer` `method`), 1256
`needs_fitting()` (`evalml.pipelines.XGBoostClassifier` `method`), 1258
`needs_fitting()` (`evalml.pipelines.XGBoostRegressor` `method`), 1260
`new()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` `method`), 1067
`new()` (`evalml.pipelines.BinaryClassificationPipeline` `method`), 1131
`new()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` `method`), 1074
`new()` (`evalml.pipelines.ClassificationPipeline` `method`), 1141
`new()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` `method`), 1082
`new()` (`evalml.pipelines.MulticlassClassificationPipeline` `method`), 1188
`new()` (`evalml.pipelines.pipeline_base.PipelineBase` `method`), 1087
`new()` (`evalml.pipelines.PipelineBase` `method`), 1198
`new()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` `method`), 1093
`new()` (`evalml.pipelines.RegressionPipeline` `method`), 1209
`new()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline` `method`), 1098
`new()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline` `method`), 1104
`new()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline` `method`), 1109
`new()` (`evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase` `method`), 1115
`new()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` `method`), 1120
`new()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` `method`), 1236
`new()` (`evalml.pipelines.TimeSeriesClassificationPipeline` `method`), 1241
`new()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` `method`), 1247
`new()` (`evalml.pipelines.TimeSeriesRegressionPipeline` `method`), 1252
`next_batch()` (`evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm` `method`), 167
`next_batch()` (`evalml.automl.automl_algorithm.AutoMLAlgorithm` `method`), 173
`next_batch()` (`evalml.automl.automl_algorithm.default_algorithm.DefaultAlgorithm` `method`), 169
`next_batch()` (`evalml.automl.automl_algorithm.DefaultAlgorithm` `method`), 174
`next_batch()` (`evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm` `method`), 171
`next_batch()` (`evalml.automl.automl_algorithm.IterativeAlgorithm` `method`), 175
`NoParamsException`, 1283, 1285
`NoPositiveLabelException`, 265, 267
`normalize_confusion_matrix()` (in module `evalml.model_understanding`), 300
`normalize_confusion_matrix()` (in module `evalml.model_understanding.graphs`), 286
`NoVarianceDataCheck` (class in `evalml.data_checks`), 254
`NoVarianceDataCheck` (class in `evalml.data_checks.no_variance_data_check`), 231
`NullsInColumnWarning`, 265, 267
`number_of_features()` (in module `evalml.preprocessing.utils`), 1271
`number_of_features()` (in module `evalml.preprocessing.utils`), 1269
`numeric_and_boolean_wv` (in module `evalml.utils.woodwork_utils`), 1294

O

`objective_function()` (`evalml.objectives.AccuracyBinary` `method`), 393
`objective_function()` (`evalml.objectives.AccuracyMulticlass` `method`), 395
`objective_function()` (`evalml.objectives.AUC` `method`), 397
`objective_function()` (`evalml.objectives.AUCMacro` `method`), 399
`objective_function()` (`evalml.objectives.AUCMicro` `method`), 401
`objective_function()` (`evalml.objectives.AUCWeighted` `method`), 402
`objective_function()` (`evalml.objectives.BalancedAccuracyBinary` `method`), 404
`objective_function()` (`evalml.objectives.BalancedAccuracyMulticlass` `method`), 406
`objective_function()`

(`evalml.objectives.binary_classification_objective.BinaryClassificationObjective`
 class method), 305
 objective_function() (`evalml.objectives.BinaryClassificationObjective`
 class method), 409
 objective_function() (`evalml.objectives.cost_benefit_matrix.CostBenefitMatrix`
 method), 308
 objective_function() (`evalml.objectives.CostBenefitMatrix` method),
 411
 objective_function() (`evalml.objectives.ExpVariance` method),
 413
 objective_function() (`evalml.objectives.F1` method), 415
 objective_function() (`evalml.objectives.F1Macro` method), 417
 objective_function() (`evalml.objectives.F1Micro` method), 418
 objective_function() (`evalml.objectives.F1Weighted` method),
 420
 objective_function() (`evalml.objectives.fraud_cost.FraudCost`
 method), 311
 objective_function() (`evalml.objectives.FraudCost` method), 422
 objective_function() (`evalml.objectives.Gini` method), 425
 objective_function() (`evalml.objectives.lead_scoring.LeadScoring`
 method), 313
 objective_function() (`evalml.objectives.LeadScoring` method),
 428
 objective_function() (`evalml.objectives.LogLossBinary` method),
 430
 objective_function() (`evalml.objectives.LogLossMulticlass` method),
 432
 objective_function() (`evalml.objectives.MAE` method), 433
 objective_function() (`evalml.objectives.MAPE` method), 435
 objective_function() (`evalml.objectives.MaxError` method), 437
 objective_function() (`evalml.objectives.MCCBinary` method),
 439
 objective_function() (`evalml.objectives.MCCMulticlass` method),
 441
 (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective`
 class method), 305
 (`evalml.objectives.MeanSquaredLogError`
 method), 442
 objective_function() (`evalml.objectives.MedianAE` method), 444
 objective_function() (`evalml.objectives.MSE`
 method), 445
 objective_function() (`evalml.objectives.multiclass_classification_objective.MulticlassC`
 class method), 316
 objective_function() (`evalml.objectives.MulticlassClassificationObjective`
 class method), 447
 objective_function() (`evalml.objectives.objective_base.ObjectiveBase`
 class method), 318
 objective_function() (`evalml.objectives.ObjectiveBase` class
 method), 450
 objective_function() (`evalml.objectives.Precision` method), 452
 objective_function() (`evalml.objectives.PrecisionMacro` method),
 454
 objective_function() (`evalml.objectives.PrecisionMicro` method),
 455
 objective_function() (`evalml.objectives.PrecisionWeighted` method),
 457
 objective_function() (`evalml.objectives.R2`
 method), 458
 objective_function() (`evalml.objectives.Recall`
 method), 460
 objective_function() (`evalml.objectives.RecallMacro` method),
 462
 objective_function() (`evalml.objectives.RecallMicro` method),
 464
 objective_function() (`evalml.objectives.RecallWeighted` method),
 465
 objective_function() (`evalml.objectives.regression_objective.RegressionObjective`
 class method), 320
 objective_function() (`evalml.objectives.RegressionObjective` class
 method), 467
 objective_function() (`evalml.objectives.RootMeanSquaredError`
 method), 469
 objective_function() (`evalml.objectives.RootMeanSquaredLogError`

<code>method)</code> , 471	<code>method)</code> , 354
<code>objective_function()</code> (<code>evalml.objectives.sensitivity_low_alert.SensitivityLowAlert</code> <code>method)</code> , 323	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MAE</code> <code>method)</code> , 355
<code>objective_function()</code> (<code>evalml.objectives.SensitivityLowAlert</code> <code>method)</code> , 473	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MAPE</code> <code>method)</code> , 357
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AccuracyBinary</code> <code>method)</code> , 326	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MaxError</code> <code>method)</code> , 359
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AccuracyMulticlass</code> <code>method)</code> , 328	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MCCBinary</code> <code>method)</code> , 361
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> <code>method)</code> , 330	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MCCMulticlass</code> <code>method)</code> , 362
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AUCMacro</code> <code>method)</code> , 332	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MeanSquaredLogError</code> <code>method)</code> , 364
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AUCMicro</code> <code>method)</code> , 334	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MedianAE</code> <code>method)</code> , 366
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.AUCWeighted</code> <code>method)</code> , 335	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.MSE</code> <code>method)</code> , 367
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.BalancedAccuracyBinary</code> <code>method)</code> , 337	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.Precision</code> <code>method)</code> , 369
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.BalancedAccuracyMulticlass</code> <code>method)</code> , 339	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.PrecisionMacro</code> <code>method)</code> , 371
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.ExpVariance</code> <code>method)</code> , 341	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.PrecisionMicro</code> <code>method)</code> , 373
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.F1</code> <code>method)</code> , 343	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.PrecisionWeighted</code> <code>method)</code> , 374
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.F1Macro</code> <code>method)</code> , 344	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.R2</code> <code>method)</code> , 376
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.F1Micro</code> <code>method)</code> , 346	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.Recall</code> <code>method)</code> , 378
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.F1Weighted</code> <code>method)</code> , 347	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.RecallMacro</code> <code>method)</code> , 380
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.Gini</code> <code>method)</code> , 349	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.RecallMicro</code> <code>method)</code> , 381
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.LogLossBinary</code> <code>method)</code> , 352	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.RecallWeighted</code> <code>method)</code> , 383
<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.LogLossMulticlass</code>	<code>objective_function()</code> (<code>evalml.objectives.standard_metrics.RootMeanSquaredError</code>

`method`), 384
`objective_function()`
 (`evalml.objectives.standard_metrics.RootMeanSquaredLogError`
 `method`), 386
`objective_function()`
 (`evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective`
 `class method`), 388
`ObjectiveBase` (`class in evalml.objectives`), 448
`ObjectiveBase` (`class in evalml.objectives.objective_base`), 317
`ObjectiveCreationError`, 265, 267
`ObjectiveNotFoundError`, 265, 267
`OneHotEncoder` (`class in evalml.pipelines`), 1190
`OneHotEncoder` (`class in evalml.pipelines.components`), 1000
`OneHotEncoder` (`class in evalml.pipelines.components.transformers`), 878
`OneHotEncoder` (`class in evalml.pipelines.components.transformers.encoders`), 715
`OneHotEncoder` (`class in evalml.pipelines.components.transformers.encoders.onehot (enable)`), 708
`OneHotEncoderMeta` (`class in evalml.pipelines.components.transformers.encoders.onehot (enable)`), 711
`optimize_threshold()`
 (`evalml.objectives.AccuracyBinary method`), 394
`optimize_threshold()` (`evalml.objectives.AUC method`), 398
`optimize_threshold()`
 (`evalml.objectives.BalancedAccuracyBinary method`), 405
`optimize_threshold()`
 (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective`
 `method`), 305
`optimize_threshold()`
 (`evalml.objectives.BinaryClassificationObjective method`), 409
`optimize_threshold()`
 (`evalml.objectives.cost_benefit_matrix.CostBenefitMatrix method`), 308
`optimize_threshold()`
 (`evalml.objectives.CostBenefitMatrix method`), 411
`optimize_threshold()` (`evalml.objectives.F1 method`), 415
`optimize_threshold()`
 (`evalml.objectives.fraud_cost.FraudCost method`), 311
`optimize_threshold()`
 (`evalml.objectives.FraudCost method`), 423
`optimize_threshold()` (`evalml.objectives.Gini method`), 425
`optimize_threshold()`
 (`evalml.objectives.lead_scoring.LeadScoring method`), 313
`optimize_threshold()`
 (`evalml.objectives.LeadScoring method`), 428
`optimize_threshold()`
 (`evalml.objectives.LogLossBinary method`), 430
`optimize_threshold()`
 (`evalml.objectives.MCCBinary method`), 439
`optimize_threshold()`
 (`evalml.objectives.Precision method`), 452
`optimize_threshold()` (`evalml.objectives.Recall method`), 461
`optimize_threshold()`
 (`evalml.objectives.sensitivity_low_alert.SensitivityLowAlert method`), 323
`optimize_threshold()`
 (`evalml.objectives.SensitivityLowAlert method`), 473
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.AccuracyBinary method`), 327
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.AUC method`), 330
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.BalancedAccuracyBinary method`), 337
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.F1 method`), 343
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.Gini method`), 350
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.LogLossBinary method`), 352
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.MCCBinary method`), 361
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.Precision method`), 369
`optimize_threshold()`
 (`evalml.objectives.standard_metrics.Recall method`), 378
`optimize_threshold()`
 (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline`
 `method`), 1067

[optimize_threshold\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipelineMixin](#) method), 1070
[optimize_threshold\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) method), 1131
[optimize_threshold\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) method), 1098
[optimize_threshold\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) method), 1236
[OutliersDataCheck](#) (class in [evalml.data_checks](#)), 255
[OutliersDataCheck](#) (class in [evalml.data_checks.outliers_data_check](#)), 232
[Oversampler](#) (class in [evalml.pipelines.components](#)), 1003
[Oversampler](#) (class in [evalml.pipelines.components.transformers](#)), 881
[Oversampler](#) (class in [evalml.pipelines.components.transformers.samplers](#)), 826
[Oversampler](#) (class in [evalml.pipelines.components.transformers.samplers.over_sampler](#)), 821
P
[pad_with_nans\(\)](#) (in module [evalml.utils](#)), 1298
[pad_with_nans\(\)](#) (in module [evalml.utils.gen_utils](#)), 1292
[ParameterError](#), 1283, 1285
[ParameterNotUsedWarning](#), 265, 267
[parameters\(\)](#) ([evalml.pipelines.ARIMARegressor](#) property), 1127
[parameters\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) property), 1068
[parameters\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) property), 1131
[parameters\(\)](#) ([evalml.pipelines.CatBoostClassifier](#) property), 1135
[parameters\(\)](#) ([evalml.pipelines.CatBoostRegressor](#) property), 1138
[parameters\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) property), 1074
[parameters\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) property), 1141
[parameters\(\)](#) ([evalml.pipelines.components.ARIMARegressor](#) property), 929
[parameters\(\)](#) ([evalml.pipelines.components.BaselineClassifier](#) property), 931
[parameters\(\)](#) ([evalml.pipelines.components.BaselineRegressor](#) property), 936
[parameters\(\)](#) ([evalml.pipelines.components.CatBoostClassifier](#) property), 936
[parameters\(\)](#) ([evalml.pipelines.components.CatBoostRegressor](#) property), 939
[parameters\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#) property), 941
[parameters\(\)](#) ([evalml.pipelines.components.ComponentBase](#) property), 941
[parameters\(\)](#) ([evalml.pipelines.components.DateTimeFeaturizer](#) property), 944
[parameters\(\)](#) ([evalml.pipelines.components.DecisionTreeClassifier](#) property), 946
[parameters\(\)](#) ([evalml.pipelines.components.DecisionTreeRegressor](#) property), 949
[parameters\(\)](#) ([evalml.pipelines.components.DelayedFeatureTransformer](#) property), 952
[parameters\(\)](#) ([evalml.pipelines.components.DFSTransformer](#) property), 954
[parameters\(\)](#) ([evalml.pipelines.components.DropColumns](#) property), 956
[parameters\(\)](#) ([evalml.pipelines.components.DropNullColumns](#) property), 958
[parameters\(\)](#) ([evalml.pipelines.components.DropRowsTransformer](#) property), 960
[parameters\(\)](#) ([evalml.pipelines.components.ElasticNetClassifier](#) property), 963
[parameters\(\)](#) ([evalml.pipelines.components.ElasticNetRegressor](#) property), 965
[parameters\(\)](#) ([evalml.pipelines.components.EmailFeaturizer](#) property), 967
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier](#) property), 476
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_regressor](#) property), 479
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier](#) property), 482
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier](#) property), 485
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier](#) property), 488
[parameters\(\)](#) ([evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor](#) property), 490
[parameters\(\)](#) ([evalml.pipelines.components.Estimator](#) property), 969
[parameters\(\)](#) ([evalml.pipelines.components.estimators.ARIMARegressor](#) property), 634
[parameters\(\)](#) ([evalml.pipelines.components.estimators.BaselineClassifier](#) property), 636
[parameters\(\)](#) ([evalml.pipelines.components.estimators.BaselineRegressor](#) property), 639
[parameters\(\)](#) ([evalml.pipelines.components.estimators.CatBoostClassifier](#) property), 641

`parameters()` (`evalml.pipelines.components.estimators.CatBoostRegressor`), 644
`parameters()` (`evalml.pipelines.components.estimators.CatBoostRegressor`), 658
`parameters()` (`evalml.pipelines.components.estimators.classifier_base`), 493
`parameters()` (`evalml.pipelines.components.estimators.classifier_base`), 630
`parameters()` (`evalml.pipelines.components.estimators.classifier_base`), 526
`parameters()` (`evalml.pipelines.components.estimators.classifier_base`), 661
`parameters()` (`evalml.pipelines.components.estimators.classifier_catboost`), 496
`parameters()` (`evalml.pipelines.components.estimators.classifier_catboost`), 664
`parameters()` (`evalml.pipelines.components.estimators.classifier_catboost`), 529
`parameters()` (`evalml.pipelines.components.estimators.classifier_catboost`), 666
`parameters()` (`evalml.pipelines.components.estimators.classifier_decision_tree`), 499
`parameters()` (`evalml.pipelines.components.estimators.classifier_decision_tree`), 669
`parameters()` (`evalml.pipelines.components.estimators.classifier_decision_tree`), 532
`parameters()` (`evalml.pipelines.components.estimators.classifier_decision_tree`), 672
`parameters()` (`evalml.pipelines.components.estimators.classifier_elasticnet`), 502
`parameters()` (`evalml.pipelines.components.estimators.classifier_elasticnet`), 674
`parameters()` (`evalml.pipelines.components.estimators.classifier_elasticnet`), 535
`parameters()` (`evalml.pipelines.components.estimators.classifier_elasticnet`), 677
`parameters()` (`evalml.pipelines.components.estimators.classifier_et_classifier`), 505
`parameters()` (`evalml.pipelines.components.estimators.classifier_et_classifier`), 680
`parameters()` (`evalml.pipelines.components.estimators.classifier_extra_trees`), 538
`parameters()` (`evalml.pipelines.components.estimators.classifier_extra_trees`), 682
`parameters()` (`evalml.pipelines.components.estimators.classifier_kneighbors`), 509
`parameters()` (`evalml.pipelines.components.estimators.classifier_kneighbors`), 684
`parameters()` (`evalml.pipelines.components.estimators.classifier_kneighbors`), 541
`parameters()` (`evalml.pipelines.components.estimators.classifier_kneighbors`), 558
`parameters()` (`evalml.pipelines.components.estimators.classifier_lightgbm`), 512
`parameters()` (`evalml.pipelines.components.estimators.classifier_lightgbm`), 596
`parameters()` (`evalml.pipelines.components.estimators.classifier_lightgbm`), 544
`parameters()` (`evalml.pipelines.components.estimators.classifier_lightgbm`), 561
`parameters()` (`evalml.pipelines.components.estimators.classifier_logistic`), 515
`parameters()` (`evalml.pipelines.components.estimators.classifier_logistic`), 599
`parameters()` (`evalml.pipelines.components.estimators.classifier_logistic`), 547
`parameters()` (`evalml.pipelines.components.estimators.classifier_logistic`), 564
`parameters()` (`evalml.pipelines.components.estimators.classifier_random_forest`), 549
`parameters()` (`evalml.pipelines.components.estimators.classifier_random_forest`), 601
`parameters()` (`evalml.pipelines.components.estimators.classifier_rf_classifier`), 518
`parameters()` (`evalml.pipelines.components.estimators.classifier_rf_classifier`), 567
`parameters()` (`evalml.pipelines.components.estimators.classifier_svm`), 521
`parameters()` (`evalml.pipelines.components.estimators.classifier_svm`), 604
`parameters()` (`evalml.pipelines.components.estimators.classifier_svm`), 552
`parameters()` (`evalml.pipelines.components.estimators.classifier_svm`), 570
`parameters()` (`evalml.pipelines.components.estimators.classifier_xgboost`), 523
`parameters()` (`evalml.pipelines.components.estimators.classifier_xgboost`), 607
`parameters()` (`evalml.pipelines.components.estimators.classifier_xgboost`), 555
`parameters()` (`evalml.pipelines.components.estimators.classifier_xgboost`), 573
`parameters()` (`evalml.pipelines.components.estimators.DecisionTreeClassifier`), 647
`parameters()` (`evalml.pipelines.components.estimators.DecisionTreeClassifier`), 610
`parameters()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor`), 650
`parameters()` (`evalml.pipelines.components.estimators.DecisionTreeRegressor`), 576
`parameters()` (`evalml.pipelines.components.estimators.ElasticNetClassifier`), 653
`parameters()` (`evalml.pipelines.components.estimators.ElasticNetClassifier`), 613
`parameters()` (`evalml.pipelines.components.estimators.ElasticNetRegressor`), 655
`parameters()` (`evalml.pipelines.components.estimators.ElasticNetRegressor`), 579

`parameters()` (`evalml.pipelines.components.estimators.regressor.LinearRegressor`), 615
`parameters()` (`evalml.pipelines.components.estimators.regressor.LinearRegressor`), 1000
`parameters()` (`evalml.pipelines.components.estimators.regressor.prophet.RegressorProphet`), 582
`parameters()` (`evalml.pipelines.components.estimators.regressor.prophet.RegressorProphet`), 1003
`parameters()` (`evalml.pipelines.components.estimators.regressor.ProphetRegressor`), 618
`parameters()` (`evalml.pipelines.components.estimators.regressor.ProphetRegressor`), 1005
`parameters()` (`evalml.pipelines.components.estimators.regressor.RandomForestRegressor`), 620
`parameters()` (`evalml.pipelines.components.estimators.regressor.RandomForestRegressor`), 1007
`parameters()` (`evalml.pipelines.components.estimators.regressor.eff_regression.RandomForestRegressor`), 585
`parameters()` (`evalml.pipelines.components.estimators.regressor.eff_regression.RandomForestRegressor`), 1009
`parameters()` (`evalml.pipelines.components.estimators.regressor.sum_regression.SVMRegressor`), 587
`parameters()` (`evalml.pipelines.components.estimators.regressor.sum_regression.SVMRegressor`), 1012
`parameters()` (`evalml.pipelines.components.estimators.regressor.SVMRegressor`), 622
`parameters()` (`evalml.pipelines.components.estimators.regressor.SVMRegressor`), 1014
`parameters()` (`evalml.pipelines.components.estimators.regressor.time_series_baseline_estimator.TimeSeriesBaselineEstimator`), 590
`parameters()` (`evalml.pipelines.components.estimators.regressor.time_series_baseline_estimator.TimeSeriesBaselineEstimator`), 1016
`parameters()` (`evalml.pipelines.components.estimators.regressor.TimeSeriesBaselineEstimator`), 625
`parameters()` (`evalml.pipelines.components.estimators.regressor.TimeSeriesBaselineEstimator`), 1019
`parameters()` (`evalml.pipelines.components.estimators.regressor.xgboost.RegressorXGBClassifier`), 593
`parameters()` (`evalml.pipelines.components.estimators.regressor.xgboost.RegressorXGBClassifier`), 1021
`parameters()` (`evalml.pipelines.components.estimators.regressor.XGBClassifier`), 627
`parameters()` (`evalml.pipelines.components.estimators.regressor.XGBClassifier`), 1024
`parameters()` (`evalml.pipelines.components.estimators.SVMClassifier`), 687
`parameters()` (`evalml.pipelines.components.estimators.SVMClassifier`), 1026
`parameters()` (`evalml.pipelines.components.estimators.SVMRegressor`), 689
`parameters()` (`evalml.pipelines.components.estimators.SVMRegressor`), 1028
`parameters()` (`evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator`), 691
`parameters()` (`evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator`), 1030
`parameters()` (`evalml.pipelines.components.estimators.XGBClassifier`), 694
`parameters()` (`evalml.pipelines.components.estimators.XGBClassifier`), 1033
`parameters()` (`evalml.pipelines.components.estimators.XGBRegressor`), 697
`parameters()` (`evalml.pipelines.components.estimators.XGBRegressor`), 1036
`parameters()` (`evalml.pipelines.components.ExtraTreesClassifier`), 972
`parameters()` (`evalml.pipelines.components.ExtraTreesClassifier`), 1038
`parameters()` (`evalml.pipelines.components.ExtraTreesRegressor`), 975
`parameters()` (`evalml.pipelines.components.ExtraTreesRegressor`), 1040
`parameters()` (`evalml.pipelines.components.FeatureSelector`), 977
`parameters()` (`evalml.pipelines.components.FeatureSelector`), 1042
`parameters()` (`evalml.pipelines.components.Imputer`), 980
`parameters()` (`evalml.pipelines.components.Imputer`), 1045
`parameters()` (`evalml.pipelines.components.KNeighborsClassifier`), 983
`parameters()` (`evalml.pipelines.components.KNeighborsClassifier`), 1047
`parameters()` (`evalml.pipelines.components.LightGBMClassifier`), 985
`parameters()` (`evalml.pipelines.components.LightGBMClassifier`), 1049
`parameters()` (`evalml.pipelines.components.LightGBMRegressor`), 988
`parameters()` (`evalml.pipelines.components.LightGBMRegressor`), 1051
`parameters()` (`evalml.pipelines.components.LinearDiscriminantAnalysis`), 990
`parameters()` (`evalml.pipelines.components.LinearDiscriminantAnalysis`), 1054
`parameters()` (`evalml.pipelines.components.LinearRegressor`), 993
`parameters()` (`evalml.pipelines.components.LinearRegressor`), 838
`parameters()` (`evalml.pipelines.components.LogisticRegressionClassifier`), 995
`parameters()` (`evalml.pipelines.components.LogisticRegressionClassifier`), 840
`parameters()` (`evalml.pipelines.components.LogTransformer`), 997
`parameters()` (`evalml.pipelines.components.LogTransformer`), 842

`parameters()` (`evalml.pipelines.components.transformers.column_selector`), 844
`parameters()` (`evalml.pipelines.components.transformers.data_time_features`), 853
`parameters()` (`evalml.pipelines.components.transformers.delayed_features`), 855
`parameters()` (`evalml.pipelines.components.transformers.deseasonalize`), 858
`parameters()` (`evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_hesanddiscriminantanalysis`), 699
`parameters()` (`evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_discriminantanalysis`), 704
`parameters()` (`evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_pca`), 706
`parameters()` (`evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_pca_and_lda`), 702
`parameters()` (`evalml.pipelines.components.transformers.drop_columns`), 860
`parameters()` (`evalml.pipelines.components.transformers.drop_null_columns`), 862
`parameters()` (`evalml.pipelines.components.transformers.drop_rows`), 864
`parameters()` (`evalml.pipelines.components.transformers.email_features`), 866
`parameters()` (`evalml.pipelines.components.transformers.encode_categorical`), 710
`parameters()` (`evalml.pipelines.components.transformers.encode_on_hot_embeddings`), 717
`parameters()` (`evalml.pipelines.components.transformers.encode_target`), 714
`parameters()` (`evalml.pipelines.components.transformers.encode_target_for_embedding`), 720
`parameters()` (`evalml.pipelines.components.transformers.feature_selector`), 722
`parameters()` (`evalml.pipelines.components.transformers.feature_selector_adaptive`), 732
`parameters()` (`evalml.pipelines.components.transformers.feature_selector_f1`), 726
`parameters()` (`evalml.pipelines.components.transformers.feature_selector_f1_micro`), 729
`parameters()` (`evalml.pipelines.components.transformers.feature_selector_f1_macro`), 735
`parameters()` (`evalml.pipelines.components.transformers.feature_selector_f1_weighted`), 737
`parameters()` (`evalml.pipelines.components.transformers.feature_selector`), 868
`parameters()` (`evalml.pipelines.components.transformers.imputer`), 871
`parameters()` (`evalml.pipelines.components.transformers.imputer_simple`), 752
`parameters()` (`evalml.pipelines.components.transformers.imputer_simple_nearest`), 740
`parameters()` (`evalml.pipelines.components.transformers.imputer_per_column`), 743

[parameters\(\) \(evalml.pipelines.pipeline_base.PipelineBase property\), 1087](#)
[parameters\(\) \(evalml.pipelines.PipelineBase property\), 1198](#)
[parameters\(\) \(evalml.pipelines.ProphetRegressor property\), 1201](#)
[parameters\(\) \(evalml.pipelines.RandomForestClassifier property\), 1203](#)
[parameters\(\) \(evalml.pipelines.RandomForestRegressor property\), 1206](#)
[parameters\(\) \(evalml.pipelines.regression_pipeline.RegressionPipeline property\), 1093](#)
[parameters\(\) \(evalml.pipelines.RegressionPipeline property\), 1209](#)
[parameters\(\) \(evalml.pipelines.RFClassifierSelectFromModel property\), 1213](#)
[parameters\(\) \(evalml.pipelines.RFRegressorSelectFromModel property\), 1215](#)
[parameters\(\) \(evalml.pipelines.SimpleImputer property\), 1217](#)
[parameters\(\) \(evalml.pipelines.SklearnStackedEnsembleClassifier property\), 1220](#)
[parameters\(\) \(evalml.pipelines.SklearnStackedEnsembleRegressor property\), 1222](#)
[parameters\(\) \(evalml.pipelines.StandardScaler property\), 1225](#)
[parameters\(\) \(evalml.pipelines.SVMClassifier property\), 1227](#)
[parameters\(\) \(evalml.pipelines.SVMRegressor property\), 1229](#)
[parameters\(\) \(evalml.pipelines.TargetEncoder property\), 1232](#)
[parameters\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1098](#)
[parameters\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1104](#)
[parameters\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property\), 1110](#)
[parameters\(\) \(evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase property\), 1115](#)
[parameters\(\) \(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline property\), 1121](#)
[parameters\(\) \(evalml.pipelines.TimeSeriesBinaryClassificationPipeline property\), 1236](#)
[parameters\(\) \(evalml.pipelines.TimeSeriesClassificationPipeline property\), 1242](#)
[parameters\(\) \(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property\), 1247](#)
[parameters\(\) \(evalml.pipelines.TimeSeriesRegressionPipeline property\), 1252](#)
[parameters\(\) \(evalml.pipelines.Transformer property\), 1256](#)
[parameters\(\) \(evalml.pipelines.XGBoostClassifier property\), 1258](#)
[parameters\(\) \(evalml.pipelines.XGBoostRegressor property\), 1261](#)
[partial_dependence\(\) \(in module evalml.model_understanding\), 300](#)
[partial_dependence\(\) \(in module evalml.model_understanding.graphs\), 286](#)
[PartialDependenceError, 265, 267](#)
[PartialDependenceErrorCode \(class in evalml.exceptions\), 267](#)
[PartialDependenceErrorCode \(class in evalml.exceptions.exceptions\), 265](#)
[PCA \(class in evalml.pipelines.components\), 1005](#)
[PCA \(class in evalml.pipelines.components.transformers\), 883](#)
[Model \(class in evalml.pipelines.components.transformers.dimensionality_reduction\), 705](#)
[Model \(class in evalml.pipelines.components.transformers.dimensionality_reduction\), 700](#)
[PerColumnImputer \(class in evalml.pipelines\), 1192](#)
[PerColumnImputer \(class in evalml.pipelines.components\), 1007](#)
[PerColumnImputer \(class in evalml.pipelines.components.transformers\), 885](#)
[PerColumnImputer \(class in evalml.pipelines.components.transformers.imputers\), 752](#)
[PerColumnImputer \(class in evalml.pipelines.components.transformers.imputers.per_column_imputer\), 741](#)
[perfect_score\(\) \(evalml.objectives.binary_classification_objective.BinaryClassificationObjective property\), 305](#)
[perfect_score\(\) \(evalml.objectives.binary_classification_objective.BinaryClassificationObjective property\), 409](#)
[perfect_score\(\) \(evalml.objectives.classification_objective.ClassificationObjective property\), 316](#)
[perfect_score\(\) \(evalml.objectives.classification_objective.MulticlassClassificationObjective property\), 448](#)
[perfect_score\(\) \(evalml.objectives.classification_objective.MulticlassClassificationObjective property\), 318](#)
[perfect_score\(\) \(evalml.objectives.classification_objective.MulticlassClassificationObjective property\), 450](#)
[perfect_score\(\) \(evalml.objectives.classification_objective.MulticlassClassificationObjective property\), 321](#)
[perfect_score\(\) \(evalml.objectives.RegressionObjective property\), 468](#)
[perfect_score\(\) \(evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective property\), 389](#)
[pipeline_number\(\) \(evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm property\), 167](#)
[pipeline_number\(\) \(evalml.automl.automl_algorithm.AutoMLAlgorithm property\), 173](#)

`pipeline_number()` (`evalml.automl.automl_algorithm.default_algorithm.DefaultAlgorithm` property), 169
`pipeline_number()` (`evalml.automl.automl_algorithm.DefaultAlgorithm` property), 174
`pipeline_number()` (`evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm` property), 171
`pipeline_number()` (`evalml.automl.automl_algorithm.IterativeAlgorithm` property), 176
`PipelineBase` (class in `evalml.pipelines`), 1195
`PipelineBase` (class in `evalml.pipelines.pipeline_base`), 1084
`PipelineBaseMeta` (class in `evalml.pipelines.pipeline_meta`), 1089
`PipelineNotFoundError`, 266, 268
`PipelineNotYetFittedError`, 266, 268
`PipelineScoreError`, 266, 268
`PipelineSearchPlots` (class in `evalml.automl.pipeline_search_plots`), 200
`plot()` (`evalml.automl.automl_search.AutoMLSearch` property), 196
`plot()` (`evalml.automl.AutoMLSearch` property), 207
`plot()` (`evalml.AutoMLSearch` property), 1303
`PolynomialDetrender` (class in `evalml.pipelines.components`), 1010
`PolynomialDetrender` (class in `evalml.pipelines.components.transformers`), 888
`PolynomialDetrender` (class in `evalml.pipelines.components.transformers.preprocessing`), 809
`PolynomialDetrender` (class in `evalml.pipelines.components.transformers.preprocessing.polynomial_detrender`), 778
`positive_only()` (`evalml.objectives.AccuracyBinary` method), 394
`positive_only()` (`evalml.objectives.AccuracyMulticlass` method), 395
`positive_only()` (`evalml.objectives.AUC` method), 398
`positive_only()` (`evalml.objectives.AUCMacro` method), 399
`positive_only()` (`evalml.objectives.AUCMicro` method), 401
`positive_only()` (`evalml.objectives.AUCWeighted` method), 402
`positive_only()` (`evalml.objectives.BalancedAccuracyBinary` method), 405
`positive_only()` (`evalml.objectives.BalancedAccuracyMulticlass` method), 406
`positive_only()` (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` method), 305
`positive_only()` (`evalml.objectives.BinaryClassificationObjective` method), 409
`positive_only()` (`evalml.objectives.cost_benefit_matrix.CostBenefitMatrix` method), 308
`positive_only()` (`evalml.objectives.CostBenefitMatrix` method), 412
`positive_only()` (`evalml.objectives.ExpVariance` method), 413
`positive_only()` (`evalml.objectives.F1` method), 415
`positive_only()` (`evalml.objectives.F1Macro` method), 417
`positive_only()` (`evalml.objectives.F1Micro` method), 419
`positive_only()` (`evalml.objectives.F1Weighted` method), 420
`positive_only()` (`evalml.objectives.fraud_cost.FraudCost` method), 311
`positive_only()` (`evalml.objectives.FraudCost` method), 423
`positive_only()` (`evalml.objectives.Gini` method), 426
`positive_only()` (`evalml.objectives.lead_scoring.LeadScoring` method), 314
`positive_only()` (`evalml.objectives.LeadScoring` method), 428
`positive_only()` (`evalml.objectives.LogLossBinary` method), 430
`positive_only()` (`evalml.objectives.LogLossMulticlass` method), 432
`positive_only()` (`evalml.objectives.MAE` method), 434
`positive_only()` (`evalml.objectives.MAPE` method), 435
`positive_only()` (`evalml.objectives.MaxError` method), 437
`positive_only()` (`evalml.objectives.MCCBinary` method), 439
`positive_only()` (`evalml.objectives.MCCMulticlass` method), 441
`positive_only()` (`evalml.objectives.MeanSquaredLogError` method), 442
`positive_only()` (`evalml.objectives.MedianAE` method), 444
`positive_only()` (`evalml.objectives.MSE` method), 446
`positive_only()` (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` method), 316
`positive_only()` (`evalml.objectives.MulticlassClassificationObjective` method), 448
`positive_only()` (`evalml.objectives.objective_base.ObjectiveBase` method), 318
`positive_only()` (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` method), 318

`method`), 450
`positive_only()` (`evalml.objectives.Precision` `method`), 452
`positive_only()` (`evalml.objectives.PrecisionMacro` `method`), 454
`positive_only()` (`evalml.objectives.PrecisionMicro` `method`), 455
`positive_only()` (`evalml.objectives.PrecisionWeighted` `method`), 457
`positive_only()` (`evalml.objectives.R2` `method`), 459
`positive_only()` (`evalml.objectives.Recall` `method`), 461
`positive_only()` (`evalml.objectives.RecallMacro` `method`), 462
`positive_only()` (`evalml.objectives.RecallMicro` `method`), 464
`positive_only()` (`evalml.objectives.RecallWeighted` `method`), 466
`positive_only()` (`evalml.objectives.regression_objective` `method`), 321
`positive_only()` (`evalml.objectives.RegressionObjective` `method`), 468
`positive_only()` (`evalml.objectives.RootMeanSquaredError` `method`), 469
`positive_only()` (`evalml.objectives.RootMeanSquaredLogError` `method`), 471
`positive_only()` (`evalml.objectives.sensitivity_low_alert` `method`), 323
`positive_only()` (`evalml.objectives.SensitivityLowAlert` `method`), 473
`positive_only()` (`evalml.objectives.standard_metrics.AccuracyBinary` `method`), 327
`positive_only()` (`evalml.objectives.standard_metrics.AccuracyMulticlass` `method`), 328
`positive_only()` (`evalml.objectives.standard_metrics.AUC` `method`), 331
`positive_only()` (`evalml.objectives.standard_metrics.AUCMacro` `method`), 332
`positive_only()` (`evalml.objectives.standard_metrics.AUCMicro` `method`), 334
`positive_only()` (`evalml.objectives.standard_metrics.AUCWeighted` `method`), 335
`positive_only()` (`evalml.objectives.standard_metrics.BalancedAccuracyBinary` `method`), 338
`positive_only()` (`evalml.objectives.standard_metrics.BalancedAccuracyMulticlass` `method`), 339
`positive_only()` (`evalml.objectives.standard_metrics.ExplainedVariance` `method`), 341
`positive_only()` (`evalml.objectives.standard_metrics.F1` `method`), 343
`positive_only()` (`evalml.objectives.standard_metrics.F1Macro` `method`), 344
`positive_only()` (`evalml.objectives.standard_metrics.F1Micro` `method`), 346
`positive_only()` (`evalml.objectives.standard_metrics.F1Weighted` `method`), 348
`positive_only()` (`evalml.objectives.standard_metrics.Gini` `method`), 350
`positive_only()` (`evalml.objectives.standard_metrics.LogLossBinary` `method`), 352
`positive_only()` (`evalml.objectives.standard_metrics.LogLossMulticlass` `method`), 354
`positive_only()` (`evalml.objectives.standard_metrics.MAE` `method`), 355
`positive_only()` (`evalml.objectives.standard_metrics.MAPE` `method`), 357
`positive_only()` (`evalml.objectives.standard_metrics.MaxError` `method`), 359
`positive_only()` (`evalml.objectives.standard_metrics.MCCBinary` `method`), 361
`positive_only()` (`evalml.objectives.standard_metrics.MCCMulticlass` `method`), 362
`positive_only()` (`evalml.objectives.standard_metrics.MeanSquaredLogError` `method`), 364
`positive_only()` (`evalml.objectives.standard_metrics.MedianAE` `method`), 366
`positive_only()` (`evalml.objectives.standard_metrics.MSE` `method`), 367
`positive_only()` (`evalml.objectives.standard_metrics.Precision` `method`), 370
`positive_only()` (`evalml.objectives.standard_metrics.PrecisionMacro` `method`), 371
`positive_only()` (`evalml.objectives.standard_metrics.PrecisionMicro` `method`), 373
`positive_only()` (`evalml.objectives.standard_metrics.PrecisionWeighted` `method`), 374
`positive_only()` (`evalml.objectives.standard_metrics.R2` `method`), 376
`positive_only()` (`evalml.objectives.standard_metrics.Recall` `method`), 378
`positive_only()` (`evalml.objectives.standard_metrics.RecallMacro` `method`), 380
`positive_only()` (`evalml.objectives.standard_metrics.RecallMicro` `method`), 381
`positive_only()` (`evalml.objectives.standard_metrics.RecallWeighted` `method`), 383
`positive_only()` (`evalml.objectives.standard_metrics.RootMeanSquaredError` `method`), 385
`positive_only()` (`evalml.objectives.standard_metrics.RootMeanSquaredLogError` `method`), 386
`positive_only()` (`evalml.objectives.time_series_regression_objective` `method`), 389
`Precision` (class in `evalml.objectives`), 450
`Precision` (class in `evalml.objectives.standard_metrics`), 368
`precision_recall_curve()` (in module `evalml.model_understanding`), 301

precision_recall_curve() (in module `evalml.model_understanding.graphs`), 287

PrecisionMacro (class in `evalml.objectives`), 453

PrecisionMacro (class in `evalml.objectives.standard_metrics`), 370

PrecisionMicro (class in `evalml.objectives`), 454

PrecisionMicro (class in `evalml.objectives.standard_metrics`), 372

PrecisionWeighted (class in `evalml.objectives`), 456

PrecisionWeighted (class in `evalml.objectives.standard_metrics`), 373

predict() (`evalml.pipelines.ARIMARegressor` method), 1127

predict() (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` method), 1068

predict() (`evalml.pipelines.BinaryClassificationPipeline` method), 1131

predict() (`evalml.pipelines.CatBoostClassifier` method), 1135

predict() (`evalml.pipelines.CatBoostRegressor` method), 1138

predict() (`evalml.pipelines.classification_pipeline.ClassificationPipeline` method), 1074

predict() (`evalml.pipelines.ClassificationPipeline` method), 1142

predict() (`evalml.pipelines.component_graph.ComponentGraph` method), 1078

predict() (`evalml.pipelines.ComponentGraph` method), 1145

predict() (`evalml.pipelines.components.ARIMARegressor` method), 929

predict() (`evalml.pipelines.components.BaselineClassifier` method), 931

predict() (`evalml.pipelines.components.BaselineRegressor` method), 934

predict() (`evalml.pipelines.components.CatBoostClassifier` method), 936

predict() (`evalml.pipelines.components.CatBoostRegressor` method), 939

predict() (`evalml.pipelines.components.DecisionTreeClassifier` method), 946

predict() (`evalml.pipelines.components.DecisionTreeRegressor` method), 949

predict() (`evalml.pipelines.components.ElasticNetClassifier` method), 963

predict() (`evalml.pipelines.components.ElasticNetRegressor` method), 965

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.BinarySklearnStackedEnsembleClassifier` method), 476

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.BinarySklearnStackedEnsembleRegressor` method), 479

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.RandomForestClassifier` method), 482

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.RandomForestRegressor` method), 485

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.SklearnStackedEnsembleClassifier` method), 488

predict() (`evalml.pipelines.components.ensemble.sklearn_stacked_ensemble.SklearnStackedEnsembleRegressor` method), 490

predict() (`evalml.pipelines.components.Estimator` method), 969

predict() (`evalml.pipelines.components.estimators.ARIMARegressor` method), 634

predict() (`evalml.pipelines.components.estimators.BaselineClassifier` method), 636

predict() (`evalml.pipelines.components.estimators.BaselineRegressor` method), 639

predict() (`evalml.pipelines.components.estimators.CatBoostClassifier` method), 641

predict() (`evalml.pipelines.components.estimators.CatBoostRegressor` method), 644

predict() (`evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier` method), 493

predict() (`evalml.pipelines.components.estimators.classifiers.BaselineClassifier` method), 526

predict() (`evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier` method), 496

predict() (`evalml.pipelines.components.estimators.classifiers.CatBoostClassifier` method), 529

predict() (`evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier` method), 499

predict() (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` method), 532

predict() (`evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.ElasticNetClassifier` method), 502

predict() (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier` method), 535

predict() (`evalml.pipelines.components.estimators.classifiers.et_classifier.ExtraTreeClassifier` method), 506

predict() (`evalml.pipelines.components.estimators.classifiers.ExtraTreeClassifier` method), 538

predict() (`evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier` method), 509

predict() (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier` method), 541

predict() (`evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier` method), 512

predict() (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier` method), 544

predict() (`evalml.pipelines.components.estimators.classifiers.logistic_regression.LogisticRegressionClassifier` method), 515

predict() (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` method), 547

predict() (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` method), 549

predict() (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` method), 518

`predict () (evalml.pipelines.components.estimators.classifiers.SVCClassifier.predict (method), 521`
`predict () (evalml.pipelines.components.estimators.classifiers.SVCClassifier.predict (method), 552`
`predict () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier.predict (method), 523`
`predict () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier.predict (method), 555`
`predict () (evalml.pipelines.components.estimators.DecisionTreeClassifier.predict (method), 647`
`predict () (evalml.pipelines.components.estimators.DecisionTreeRegressor.predict (method), 650`
`predict () (evalml.pipelines.components.estimators.ElasticNetClassifier.predict (method), 653`
`predict () (evalml.pipelines.components.estimators.ElasticNetRegressor.predict (method), 656`
`predict () (evalml.pipelines.components.estimators.Estimator.predict (method), 658`
`predict () (evalml.pipelines.components.estimators.estimator.Estimator.predict (method), 630`
`predict () (evalml.pipelines.components.estimators.ExtraTreesClassifier.predict (method), 661`
`predict () (evalml.pipelines.components.estimators.ExtraTreesRegressor.predict (method), 664`
`predict () (evalml.pipelines.components.estimators.KNeighborsClassifier.predict (method), 666`
`predict () (evalml.pipelines.components.estimators.LightGBMClassifier.predict (method), 669`
`predict () (evalml.pipelines.components.estimators.LightGBMRegressor.predict (method), 672`
`predict () (evalml.pipelines.components.estimators.LinearRegressor.predict (method), 675`
`predict () (evalml.pipelines.components.estimators.LogisticRegressionClassifier.predict (method), 677`
`predict () (evalml.pipelines.components.estimators.ProphetRegressor.predict (method), 680`
`predict () (evalml.pipelines.components.estimators.RandomForestClassifier.predict (method), 682`
`predict () (evalml.pipelines.components.estimators.RandomForestRegressor.predict (method), 684`
`predict () (evalml.pipelines.components.estimators.regressors.Arimo.predict (method), 558`
`predict () (evalml.pipelines.components.estimators.regressors.Arimo.predict (method), 596`
`predict () (evalml.pipelines.components.estimators.regressors.baseline.BaselineRegressor.predict (method), 561`
`predict () (evalml.pipelines.components.estimators.regressors.BaselineRegressor.predict (method), 599`
`predict () (evalml.pipelines.components.estimators.regressors.CatBoostRegressor.predict (method), 564`
`predict () (evalml.pipelines.components.estimators.regressors.CatBoostRegressor.predict (method), 601`
`predict () (evalml.pipelines.components.estimators.regressors.decision.DecisionTreeRegressor.predict (method), 567`

[predict \(\) \(evalml.pipelines.components.LightGBMClassifier method\), 1184](#)
[method\), 985](#)
[predict \(\) \(evalml.pipelines.components.LightGBMRegressor method\), 1082](#)
[method\), 988](#)
[predict \(\) \(evalml.pipelines.components.LinearRegressor method\), 1188](#)
[method\), 993](#)
[predict \(\) \(evalml.pipelines.components.LogisticRegressionClassifier method\), 1087](#)
[method\), 995](#)
[predict \(\) \(evalml.pipelines.components.ProphetRegressor method\), 1198](#)
[method\), 1014](#)
[predict \(\) \(evalml.pipelines.components.RandomForestClassifier method\), 1201](#)
[method\), 1016](#)
[predict \(\) \(evalml.pipelines.components.RandomForestRegressor method\), 1203](#)
[method\), 1019](#)
[predict \(\) \(evalml.pipelines.components.SklearnStackedEnsembleClassifier method\), 1206](#)
[method\), 1033](#)
[predict \(\) \(evalml.pipelines.components.SklearnStackedEnsembleRegressor method\), 1093](#)
[method\), 1036](#)
[predict \(\) \(evalml.pipelines.components.SVMClassifier method\), 1209](#)
[method\), 1040](#)
[predict \(\) \(evalml.pipelines.components.SVMRegressor method\), 1220](#)
[method\), 1042](#)
[predict \(\) \(evalml.pipelines.components.TimeSeriesBaselineEstimator method\), 1222](#)
[method\), 1051](#)
[predict \(\) \(evalml.pipelines.components.utils.WrappedSKClassifier method\), 1227](#)
[method\), 923](#)
[predict \(\) \(evalml.pipelines.components.utils.WrappedSKRegressor method\), 1229](#)
[method\), 924](#)
[predict \(\) \(evalml.pipelines.components.XGBoostClassifier method\), 1098](#)
[method\), 1061](#)
[predict \(\) \(evalml.pipelines.components.XGBoostRegressor method\), 1104](#)
[method\), 1063](#)
[predict \(\) \(evalml.pipelines.DecisionTreeClassifier method\), 1110](#)
[method\), 1148](#)
[predict \(\) \(evalml.pipelines.DecisionTreeRegressor method\), 1116](#)
[method\), 1151](#)
[predict \(\) \(evalml.pipelines.ElasticNetClassifier method\), 1121](#)
[method\), 1158](#)
[predict \(\) \(evalml.pipelines.ElasticNetRegressor method\), 1236](#)
[method\), 1161](#)
[predict \(\) \(evalml.pipelines.Estimator method\), 1163](#)
[predict \(\) \(evalml.pipelines.ExtraTreesClassifier method\), 1247](#)
[method\), 1166](#)
[predict \(\) \(evalml.pipelines.ExtraTreesRegressor method\), 1253](#)
[method\), 1169](#)
[predict \(\) \(evalml.pipelines.KNeighborsClassifier method\), 1258](#)
[method\), 1174](#)
[predict \(\) \(evalml.pipelines.LightGBMClassifier method\), 1261](#)
[method\), 1177](#)
[predict \(\) \(evalml.pipelines.LightGBMRegressor method\), 1261](#)
[method\), 1179](#)
[predict \(\) \(evalml.pipelines.LinearRegressor method\), 1261](#)
[method\), 1182](#)
[predict \(\) \(evalml.pipelines.LogisticRegressionClassifier method\), 1261](#)

method), 1104

`predict_in_sample()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* *method*), 1110

`predict_in_sample()` (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase* *method*), 1116

`predict_in_sample()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* *method*), 1121

`predict_in_sample()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* *method*), 1236

`predict_in_sample()` (*evalml.pipelines.TimeSeriesClassificationPipeline* *method*), 1242

`predict_in_sample()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* *method*), 1248

`predict_in_sample()` (*evalml.pipelines.TimeSeriesRegressionPipeline* *method*), 1253

`predict_proba()` (*evalml.pipelines.ARIMAREgressor* *method*), 1127

`predict_proba()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* *method*), 1068

`predict_proba()` (*evalml.pipelines.BinaryClassificationPipeline* *method*), 1132

`predict_proba()` (*evalml.pipelines.CatBoostClassifier* *method*), 1135

`predict_proba()` (*evalml.pipelines.CatBoostRegressor* *method*), 1138

`predict_proba()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* *method*), 1074

`predict_proba()` (*evalml.pipelines.ClassificationPipeline* *method*), 1142

`predict_proba()` (*evalml.pipelines.components.ARIMAREgressor* *method*), 929

`predict_proba()` (*evalml.pipelines.components.BaselineClassifier* *method*), 931

`predict_proba()` (*evalml.pipelines.components.BaselineRegressor* *method*), 934

`predict_proba()` (*evalml.pipelines.components.CatBoostClassifier* *method*), 936

`predict_proba()` (*evalml.pipelines.components.CatBoostRegressor* *method*), 939

`predict_proba()` (*evalml.pipelines.components.DecisionTreeClassifier* *method*), 946

`predict_proba()` (*evalml.pipelines.components.DecisionTreeRegressor* *method*), 949

`predict_proba()` (*evalml.pipelines.components.ElasticNetClassifier* *method*), 963

`predict_proba()` (*evalml.pipelines.components.ElasticNetRegressor* *method*), 965

`predict_proba()` (*evalml.pipelines.components.ensemble.sklearn_stack_ensemble.StackedEnsembleClassifier* *method*), 476

`predict_proba()` (*evalml.pipelines.components.ensemble.sklearn_stack_ensemble.StackedEnsembleRegressor* *method*), 479

`predict_proba()` (*evalml.pipelines.components.ensemble.sklearn_stack_ensemble.StackedEnsembleClassifier* *method*), 482

`predict_proba()` (*evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier* *method*), 485

`predict_proba()` (*evalml.pipelines.components.ensemble.SklearnStackedEnsembleRegressor* *method*), 488

`predict_proba()` (*evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier* *method*), 490

`predict_proba()` (*evalml.pipelines.components.Estimator* *method*), 969

`predict_proba()` (*evalml.pipelines.components.estimators.ARIMAREgressor* *method*), 634

`predict_proba()` (*evalml.pipelines.components.estimators.BaselineClassifier* *method*), 636

`predict_proba()` (*evalml.pipelines.components.estimators.BaselineRegressor* *method*), 639

`predict_proba()` (*evalml.pipelines.components.estimators.CatBoostClassifier* *method*), 641

`predict_proba()` (*evalml.pipelines.components.estimators.CatBoostRegressor* *method*), 644

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.BaselineClassifier* *method*), 493

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.BaselineClassifier* *method*), 526

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 496

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 529

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 499

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 532

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 502

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 535

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 506

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 538

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 509

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 541

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 512

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* *method*), 544

`predict_proba()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* *method*), 515

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier`), 547

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier`), 550

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.RidgeClassifier`), 518

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.SVCClassifier`), 521

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier`), 552

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier`), 523

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.XGBClassifier`), 555

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier`), 647

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeRegressor`), 650

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 653

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 656

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 658

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 630

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 661

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 664

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 666

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 669

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 672

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 675

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 677

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 680

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 682

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 684

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 558

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 596

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 561

`predict_proba()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 599

`predict_proba()` (`evalml.pipelines.components.ExtraTreesClassifier` method), 972
`predict_proba()` (`evalml.pipelines.components.ExtraTreesRegressor` method), 975
`predict_proba()` (`evalml.pipelines.components.KNeighborsClassifier` method), 983
`predict_proba()` (`evalml.pipelines.components.LightGBMClassifier` method), 985
`predict_proba()` (`evalml.pipelines.components.LightGBMRegressor` method), 988
`predict_proba()` (`evalml.pipelines.components.LinearRegressor` method), 993
`predict_proba()` (`evalml.pipelines.components.LogisticRegressionClassifier` method), 995
`predict_proba()` (`evalml.pipelines.components.ProphetRegressor` method), 1014
`predict_proba()` (`evalml.pipelines.components.RandomForestClassifier` method), 1017
`predict_proba()` (`evalml.pipelines.components.RandomForestRegressor` method), 1019
`predict_proba()` (`evalml.pipelines.components.SklearnStackedEnsembleClassifier` method), 1033
`predict_proba()` (`evalml.pipelines.components.SklearnStackedEnsembleRegressor` method), 1036
`predict_proba()` (`evalml.pipelines.components.SVMClassifier` method), 1040
`predict_proba()` (`evalml.pipelines.components.SVMRegressor` method), 1042
`predict_proba()` (`evalml.pipelines.components.TimeSeriesBaselineEstimator` method), 1052
`predict_proba()` (`evalml.pipelines.components.utils.WrappedSKClassifier` method), 923
`predict_proba()` (`evalml.pipelines.components.XGBoostClassifier` method), 1061
`predict_proba()` (`evalml.pipelines.components.XGBoostRegressor` method), 1063
`predict_proba()` (`evalml.pipelines.DecisionTreeClassifier` method), 1148
`predict_proba()` (`evalml.pipelines.DecisionTreeRegressor` method), 1151
`predict_proba()` (`evalml.pipelines.ElasticNetClassifier` method), 1159
`predict_proba()` (`evalml.pipelines.ElasticNetRegressor` method), 1161
`predict_proba()` (`evalml.pipelines.Estimator` method), 1163
`predict_proba()` (`evalml.pipelines.ExtraTreesClassifier` method), 1166
`predict_proba()` (`evalml.pipelines.ExtraTreesRegressor` method), 1169
`predict_proba()` (`evalml.pipelines.KNeighborsClassifier` method), 1174
`predict_proba()` (`evalml.pipelines.LightGBMClassifier` method), 1177
`predict_proba()` (`evalml.pipelines.LightGBMRegressor` method), 1180
`predict_proba()` (`evalml.pipelines.LinearRegressor` method), 1182
`predict_proba()` (`evalml.pipelines.LogisticRegressionClassifier` method), 1184
`predict_proba()` (`evalml.pipelines.multiclass_classification_pipeline` method), 1083
`predict_proba()` (`evalml.pipelines.MulticlassClassificationPipeline` method), 1189
`predict_proba()` (`evalml.pipelines.ProphetRegressor` method), 1201
`predict_proba()` (`evalml.pipelines.RandomForestClassifier` method), 1203
`predict_proba()` (`evalml.pipelines.RandomForestRegressor` method), 1206
`predict_proba()` (`evalml.pipelines.SklearnStackedEnsembleClassifier` method), 1220
`predict_proba()` (`evalml.pipelines.SklearnStackedEnsembleRegressor` method), 1223
`predict_proba()` (`evalml.pipelines.SVMClassifier` method), 1227
`predict_proba()` (`evalml.pipelines.SVMRegressor` method), 1229
`predict_proba()` (`evalml.pipelines.time_series_classification_pipeline` method), 1099
`predict_proba()` (`evalml.pipelines.time_series_classification_pipeline` method), 1105
`predict_proba()` (`evalml.pipelines.time_series_classification_pipeline` method), 1110
`predict_proba()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` method), 1237
`predict_proba()` (`evalml.pipelines.TimeSeriesClassificationPipeline` method), 1242
`predict_proba()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` method), 1248
`predict_proba()` (`evalml.pipelines.XGBoostClassifier` method), 1258
`predict_proba()` (`evalml.pipelines.XGBoostRegressor` method), 1261
`predict_proba_in_sample()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeries` method), 1099
`predict_proba_in_sample()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeries` method), 1105
`predict_proba_in_sample()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeries` method), 1111
`predict_proba_in_sample()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` method), 1237
`predict_proba_in_sample()` (`evalml.pipelines.TimeSeriesClassificationPipeline` method), 1237

method), 1243

predict_proba_in_sample() (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method), 1248

print_deps() (in module evalml.utils.cli_utils), 1289

print_info() (in module evalml.utils.cli_utils), 1289

print_sys_info() (in module evalml.utils.cli_utils), 1289

ProblemTypes (class in evalml.problem_types), 1277

ProblemTypes (class in evalml.problem_types.problem_types), 1273

ProphetRegressor (class in evalml.pipelines), 1199

ProphetRegressor (class in evalml.pipelines.components), 1012

ProphetRegressor (class in evalml.pipelines.components.estimators), 678

ProphetRegressor (class in evalml.pipelines.components.estimators.regressors), 615

ProphetRegressor (class in evalml.pipelines.components.estimators.regressors.prophet_regressor), 580

propose() (evalml.tuners.grid_search_tuner.GridSearchTuner method), 1279

propose() (evalml.tuners.GridSearchTuner method), 1284

propose() (evalml.tuners.random_search_tuner.RandomSearchTuner method), 1280

propose() (evalml.tuners.RandomSearchTuner method), 1286

propose() (evalml.tuners.skopt_tuner.SKOptTuner method), 1282

propose() (evalml.tuners.SKOptTuner method), 1286

propose() (evalml.tuners.Tuner method), 1287

propose() (evalml.tuners.tuner.Tuner method), 1283

R

R2 (class in evalml.objectives), 458

R2 (class in evalml.objectives.standard_metrics), 375

raise_error_callback() (in module evalml.automl.callbacks), 199

RandomForestClassifier (class in evalml.pipelines), 1202

RandomForestClassifier (class in evalml.pipelines.components), 1015

RandomForestClassifier (class in evalml.pipelines.components.estimators), 680

RandomForestClassifier (class in evalml.pipelines.components.estimators.classifiers), 547

RandomForestClassifier (class in evalml.pipelines.components.estimators.classifiers.rf_classifier), 516

RandomForestRegressor (class in evalml.pipelines), 1204

RandomForestRegressor (class in evalml.pipelines.components), 1017

RandomForestRegressor (class in evalml.pipelines.components.estimators), 682

RandomForestRegressor (class in evalml.pipelines.components.estimators.regressors), 618

RandomForestRegressor (class in evalml.pipelines.components.estimators.regressors.rf_regressor), 583

RandomSearchTuner (class in evalml.tuners), 1285

RandomSearchTuner (class in evalml.tuners.random_search_tuner), 1279

rankings() (evalml.automl.automl_search.AutoMLSearch property), 196

rankings() (evalml.automl.AutoMLSearch property), 207

rankings() (evalml.automl.AutoMLSearch property), 1303

Recall (class in evalml.objectives), 459

Recall (class in evalml.objectives.standard_metrics), 376

RecallMacro (class in evalml.objectives), 461

RecallMacro (class in evalml.objectives.standard_metrics), 379

RecallMicro (class in evalml.objectives), 463

RecallMicro (class in evalml.objectives.standard_metrics), 380

RecallWeighted (class in evalml.objectives), 465

RecallWeighted (class in evalml.objectives.standard_metrics), 382

register() (evalml.pipelines.components.component_base_meta.ComponentBaseMeta method), 920

register() (evalml.pipelines.components.ComponentBaseMeta method), 941

register() (evalml.pipelines.components.transformers.encoders.onehot_encoder method), 711

register() (evalml.pipelines.components.transformers.imputers.target_imputer method), 749

register() (evalml.pipelines.pipeline_meta.PipelineBaseMeta method), 1089

register() (evalml.utils.base_meta.BaseMeta method), 1288

RegressionObjective (class in evalml.objectives), 466

RegressionObjective (class in evalml.objectives.regression_objective), 319

RegressionPipeline (class in evalml.pipelines), 1206

RegressionPipeline (class in evalml.pipelines.regression_pipeline), 1090

[results\(\)](#) ([evalml.automl.automl_search.AutoMLSearch](#) property), 196
[results\(\)](#) ([evalml.AutoMLSearch](#) property), 207
[results\(\)](#) ([evalml.AutoMLSearch](#) property), 1303
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines](#)), 1210
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components](#)), 1019
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers](#)), 890
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection](#)), 732
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection.feature_selector](#)), 723
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines](#)), 1213
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components](#)), 1022
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers](#)), 893
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection](#)), 735
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection.feature_selector](#)), 727
[roc_curve\(\)](#) (in module [evalml.model_understanding](#)), 302
[roc_curve\(\)](#) (in module [evalml.model_understanding.graphs](#)), 288
[RootMeanSquaredError](#) (class in [evalml.objectives](#)), 468
[RootMeanSquaredError](#) (class in [evalml.objectives.standard_metrics](#)), 384
[RootMeanSquaredLogError](#) (class in [evalml.objectives](#)), 470
[RootMeanSquaredLogError](#) (class in [evalml.objectives.standard_metrics](#)), 385
S
[safe_repr\(\)](#) (in module [evalml.utils](#)), 1298
[safe_repr\(\)](#) (in module [evalml.utils.gen_utils](#)), 1292
[SamplerBase](#) (class in [evalml.preprocessing.data_splitters](#)), 1267
[SamplerBase](#) (class in [evalml.preprocessing.data_splitters.sampler_base](#)), 1263
[save\(\)](#) ([evalml.automl.automl_search.AutoMLSearch](#) method), 196
[save\(\)](#) ([evalml.AutoMLSearch](#) method), 1303
[save\(\)](#) ([evalml.pipelines.ARIMARegressor](#) method), 1127
[save\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) method), 1068
[save\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) method), 1132
[save\(\)](#) ([evalml.pipelines.CatBoostClassifier](#) method), 1135
[save\(\)](#) ([evalml.pipelines.CatBoostRegressor](#) method), 1138
[save\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) method), 1074
[save\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) method), 1142
[save\(\)](#) ([evalml.pipelines.classification_pipeline.feature_selector.ARIMARegressor](#) method), 929
[save\(\)](#) ([evalml.pipelines.components.BaselineClassifier](#) method), 932
[save\(\)](#) ([evalml.pipelines.components.BaselineRegressor](#) method), 934
[save\(\)](#) ([evalml.pipelines.components.CatBoostClassifier](#) method), 936
[save\(\)](#) ([evalml.pipelines.components.CatBoostRegressor](#) method), 939
[save\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#) method), 920
[save\(\)](#) ([evalml.pipelines.components.ComponentBase](#) method), 941
[save\(\)](#) ([evalml.pipelines.components.DateFeatureSelector](#) method), 941
[save\(\)](#) ([evalml.pipelines.components.DateFeatureSelector](#) method), 944
[save\(\)](#) ([evalml.pipelines.components.DecisionTreeClassifier](#) method), 947
[save\(\)](#) ([evalml.pipelines.components.DecisionTreeRegressor](#) method), 949
[save\(\)](#) ([evalml.pipelines.components.DelayedFeatureTransformer](#) method), 952
[save\(\)](#) ([evalml.pipelines.components.DFSTransformer](#) method), 954
[save\(\)](#) ([evalml.pipelines.components.DropColumns](#) method), 956
[save\(\)](#) ([evalml.pipelines.components.DropNullColumns](#) method), 958
[save\(\)](#) ([evalml.pipelines.components.DropRowsTransformer](#) method), 960
[save\(\)](#) ([evalml.pipelines.components.ElasticNetClassifier](#) method), 963
[save\(\)](#) ([evalml.pipelines.components.ElasticNetRegressor](#) method), 965
[save\(\)](#) ([evalml.pipelines.components.EmailFeaturizer](#) method), 967
[save\(\)](#) ([evalml.pipelines.components.ensemble.sklearn_stacked_ensemble](#) method), 477

[save \(\) \(evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.RandomForestClassifier method\), 480](#)
[save \(\) \(evalml.pipelines.components.ensemble.sklearn_stacked_ensemble_classifier.RandomForestRegressor method\), 482](#)
[save \(\) \(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier method\), 485](#)
[save \(\) \(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier method\), 488](#)
[save \(\) \(evalml.pipelines.components.ensemble.SklearnStackedEnsembleClassifier method\), 491](#)
[save \(\) \(evalml.pipelines.components.Estimator method\), 970](#)
[save \(\) \(evalml.pipelines.components.estimators.ARIMARegressor method\), 634](#)
[save \(\) \(evalml.pipelines.components.estimators.BaselineClassifier method\), 637](#)
[save \(\) \(evalml.pipelines.components.estimators.BaselineRegressor method\), 639](#)
[save \(\) \(evalml.pipelines.components.estimators.CatBoostClassifier method\), 642](#)
[save \(\) \(evalml.pipelines.components.estimators.CatBoostRegressor method\), 644](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier method\), 493](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier method\), 526](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier method\), 496](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier method\), 529](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.decision_tree_classifier.DecisionTreeClassifier method\), 499](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.decision_tree_classifier.DecisionTreeClassifier method\), 532](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.elastic_net_classifier.ElasticNetClassifier method\), 503](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.elastic_net_classifier.ElasticNetClassifier method\), 535](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.extra_trees_classifier.ExtraTreesClassifier method\), 506](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.extra_trees_classifier.ExtraTreesClassifier method\), 538](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier method\), 509](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier method\), 541](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier method\), 512](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier method\), 544](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier.LogisticRegressionClassifier method\), 515](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier.LogisticRegressionClassifier method\), 547](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.sklearn_stacked_ensemble_classifier.RandomForestClassifier method\), 550](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.sklearn_stacked_ensemble_classifier.RandomForestRegressor method\), 518](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.svm_classifier.SVMClassifier method\), 521](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.SVMClassifier method\), 552](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.xgboost_classifier.XGBoostClassifier method\), 524](#)
[save \(\) \(evalml.pipelines.components.estimators.classifiers.XGBoostClassifier method\), 555](#)
[save \(\) \(evalml.pipelines.components.estimators.DecisionTreeClassifier method\), 647](#)
[save \(\) \(evalml.pipelines.components.estimators.DecisionTreeRegressor method\), 650](#)
[save \(\) \(evalml.pipelines.components.estimators.ElasticNetClassifier method\), 653](#)
[save \(\) \(evalml.pipelines.components.estimators.ElasticNetRegressor method\), 656](#)
[save \(\) \(evalml.pipelines.components.estimators.Estimator method\), 658](#)
[save \(\) \(evalml.pipelines.components.estimators.estimator.Estimator method\), 630](#)
[save \(\) \(evalml.pipelines.components.estimators.ExtraTreesClassifier method\), 661](#)
[save \(\) \(evalml.pipelines.components.estimators.ExtraTreesRegressor method\), 664](#)
[save \(\) \(evalml.pipelines.components.estimators.KNeighborsClassifier method\), 667](#)
[save \(\) \(evalml.pipelines.components.estimators.LightGBMClassifier method\), 670](#)
[save \(\) \(evalml.pipelines.components.estimators.LightGBMRegressor method\), 672](#)
[save \(\) \(evalml.pipelines.components.estimators.LinearRegressor method\), 675](#)
[save \(\) \(evalml.pipelines.components.estimators.LogisticRegressionClassifier method\), 677](#)
[save \(\) \(evalml.pipelines.components.estimators.ProphetRegressor method\), 680](#)
[save \(\) \(evalml.pipelines.components.estimators.RandomForestClassifier method\), 682](#)
[save \(\) \(evalml.pipelines.components.estimators.RandomForestRegressor method\), 685](#)
[save \(\) \(evalml.pipelines.components.estimators.regressors.arima_regressor.ARIMARegressor method\), 558](#)
[save \(\) \(evalml.pipelines.components.estimators.regressors.ARIMARegressor method\), 597](#)
[save \(\) \(evalml.pipelines.components.estimators.regressors.baseline_regressor.BaselineRegressor method\), 561](#)
[save \(\) \(evalml.pipelines.components.estimators.regressors.logistic_regression_classifier.LogisticRegressionClassifier method\), 599](#)
[save \(\) \(evalml.pipelines.components.estimators.regressors.catboost_regressor.CatBoostRegressor method\), 564](#)

`save()` (`evalml.pipelines.components.estimators.regressors.GradientBoostingRegressor` method), 602
`save()` (`evalml.pipelines.components.estimators.regressors.decision_tree.DecisionTreeRegressor` method), 567
`save()` (`evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor` method), 604
`save()` (`evalml.pipelines.components.estimators.regressors.elasticnet.ElasticNetRegressor` method), 570
`save()` (`evalml.pipelines.components.estimators.regressors.ElasticNetRegressor` method), 607
`save()` (`evalml.pipelines.components.estimators.regressors.linear.LinearRegression` method), 573
`save()` (`evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor` method), 610
`save()` (`evalml.pipelines.components.estimators.regressors.lightgbm.LightGBMRegressor` method), 576
`save()` (`evalml.pipelines.components.estimators.regressors.LightGBMRegressor` method), 613
`save()` (`evalml.pipelines.components.estimators.regressors.linear.LinearRegression` method), 579
`save()` (`evalml.pipelines.components.estimators.regressors.LinearRegression` method), 615
`save()` (`evalml.pipelines.components.estimators.regressors.prophet.ProphetRegressor` method), 582
`save()` (`evalml.pipelines.components.estimators.regressors.ProphetRegressor` method), 618
`save()` (`evalml.pipelines.components.estimators.regressors.RandomForestRegressor` method), 620
`save()` (`evalml.pipelines.components.estimators.regressors.sage.SageRegressor` method), 585
`save()` (`evalml.pipelines.components.estimators.regressors.svm.SVRRegressor` method), 588
`save()` (`evalml.pipelines.components.estimators.regressors.SVRRegressor` method), 623
`save()` (`evalml.pipelines.components.estimators.regressors.time_series.TimeSeriesRegressor` method), 590
`save()` (`evalml.pipelines.components.estimators.regressors.TimeSeriesRegressor` method), 625
`save()` (`evalml.pipelines.components.estimators.regressors.xgboost.XGBRegressor` method), 593
`save()` (`evalml.pipelines.components.estimators.regressors.XGBRegressor` method), 627
`save()` (`evalml.pipelines.components.estimators.SVMClassifier` method), 687
`save()` (`evalml.pipelines.components.estimators.SVMRegressor` method), 689
`save()` (`evalml.pipelines.components.estimators.TimeSeriesBaseLineEstimator` method), 692
`save()` (`evalml.pipelines.components.estimators.XGBoostClassifier` method), 694
`save()` (`evalml.pipelines.components.estimators.XGBoostRegressor` method), 697
`save()` (`evalml.pipelines.components.ExtraTreesClassifier` method), 972
`save()` (`evalml.pipelines.components.ExtraTreesRegressor` method), 975
`save()` (`evalml.pipelines.components.FeatureSelector` method), 977
`save()` (`evalml.pipelines.components.Imputer` method), 980
`save()` (`evalml.pipelines.components.KNeighborsClassifier` method), 983
`save()` (`evalml.pipelines.components.LightGBMClassifier` method), 986
`save()` (`evalml.pipelines.components.LightGBMRegressor` method), 988
`save()` (`evalml.pipelines.components.LinearDiscriminantAnalysis` method), 990
`save()` (`evalml.pipelines.components.LinearRegression` method), 993
`save()` (`evalml.pipelines.components.LogisticRegressionClassifier` method), 996
`save()` (`evalml.pipelines.components.LogTransformer` method), 998
`save()` (`evalml.pipelines.components.LSA` method), 1000
`save()` (`evalml.pipelines.components.OneHotEncoder` method), 1003
`save()` (`evalml.pipelines.components.Oversampler` method), 1005
`save()` (`evalml.pipelines.components.PCA` method), 1007
`save()` (`evalml.pipelines.components.PerColumnImputer` method), 1009
`save()` (`evalml.pipelines.components.PolynomialDetrender` method), 1012
`save()` (`evalml.pipelines.components.ProphetRegressor` method), 1014
`save()` (`evalml.pipelines.components.RandomForestClassifier` method), 1017
`save()` (`evalml.pipelines.components.RandomForestRegressor` method), 1019
`save()` (`evalml.pipelines.components.RFClassifierSelectFromModel` method), 1021
`save()` (`evalml.pipelines.components.RFRegressorSelectFromModel` method), 1024
`save()` (`evalml.pipelines.components.SelectByType` method), 1026
`save()` (`evalml.pipelines.components.SelectColumns` method), 1028
`save()` (`evalml.pipelines.components.SimpleImputer` method), 1030
`save()` (`evalml.pipelines.components.SklearnStackedEnsembleClassifier` method), 1033
`save()` (`evalml.pipelines.components.SklearnStackedEnsembleRegressor` method), 1036
`save()` (`evalml.pipelines.components.StandardScaler` method), 1038

<code>save()</code> (<code>evalml.pipelines.components.SVMClassifier</code> method), 1040	<code>save()</code> (<code>evalml.pipelines.components.transformers.feature_selection.FeatureSelector</code> method), 732
<code>save()</code> (<code>evalml.pipelines.components.SVMRegressor</code> method), 1043	<code>save()</code> (<code>evalml.pipelines.components.transformers.feature_selection.rf_classifier</code> method), 726
<code>save()</code> (<code>evalml.pipelines.components.TargetEncoder</code> method), 1045	<code>save()</code> (<code>evalml.pipelines.components.transformers.feature_selection.rf_regressor</code> method), 729
<code>save()</code> (<code>evalml.pipelines.components.TargetImputer</code> method), 1047	<code>save()</code> (<code>evalml.pipelines.components.transformers.feature_selection.RFCFeatureSelector</code> method), 735
<code>save()</code> (<code>evalml.pipelines.components.TextFeaturizer</code> method), 1049	<code>save()</code> (<code>evalml.pipelines.components.transformers.feature_selection.RFRFeatureSelector</code> method), 737
<code>save()</code> (<code>evalml.pipelines.components.TimeSeriesBaselineEstimator</code> method), 1052	<code>save()</code> (<code>evalml.pipelines.components.transformers.FeatureSelector</code> method), 868
<code>save()</code> (<code>evalml.pipelines.components.Transformer</code> method), 1054	<code>save()</code> (<code>evalml.pipelines.components.transformers.Imputer</code> method), 871
<code>save()</code> (<code>evalml.pipelines.components.transformers.columns_selector.CaliforniaSelector</code> method), 838	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.Imputer</code> method), 752
<code>save()</code> (<code>evalml.pipelines.components.transformers.columns_selector.DropColumns</code> method), 840	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.imputer.Imputer</code> method), 740
<code>save()</code> (<code>evalml.pipelines.components.transformers.columns_selector.DropByType</code> method), 842	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.per_column.Imputer</code> method), 743
<code>save()</code> (<code>evalml.pipelines.components.transformers.columns_selector.ScaledByType</code> method), 844	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.PerColumnImputer</code> method), 754
<code>save()</code> (<code>evalml.pipelines.components.transformers.DateTimeFeaturizer</code> method), 853	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.simple_imputer.Imputer</code> method), 746
<code>save()</code> (<code>evalml.pipelines.components.transformers.DelayedFeatureTransformer</code> method), 855	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.SimpleImputer</code> method), 757
<code>save()</code> (<code>evalml.pipelines.components.transformers.DFTransformer</code> method), 858	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.target_imputer.Imputer</code> method), 748
<code>save()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reduction.DiscretizationTransformer</code> method), 699	<code>save()</code> (<code>evalml.pipelines.components.transformers.imputers.TargetImputer</code> method), 759
<code>save()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reduction.DiscretizationTransformer</code> method), 704	<code>save()</code> (<code>evalml.pipelines.components.transformers.LinearDiscriminantAnalysis</code> method), 873
<code>save()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reduction.PCA</code> method), 707	<code>save()</code> (<code>evalml.pipelines.components.transformers.LogTransformer</code> method), 875
<code>save()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reduction.PCA</code> method), 702	<code>save()</code> (<code>evalml.pipelines.components.transformers.LSA</code> method), 877
<code>save()</code> (<code>evalml.pipelines.components.transformers.DropColumns</code> method), 860	<code>save()</code> (<code>evalml.pipelines.components.transformers.OneHotEncoder</code> method), 880
<code>save()</code> (<code>evalml.pipelines.components.transformers.DropNAColumns</code> method), 862	<code>save()</code> (<code>evalml.pipelines.components.transformers.Oversampler</code> method), 883
<code>save()</code> (<code>evalml.pipelines.components.transformers.DropRowsTransformer</code> method), 864	<code>save()</code> (<code>evalml.pipelines.components.transformers.PCA</code> method), 885
<code>save()</code> (<code>evalml.pipelines.components.transformers.EmailFeaturizer</code> method), 866	<code>save()</code> (<code>evalml.pipelines.components.transformers.PerColumnImputer</code> method), 887
<code>save()</code> (<code>evalml.pipelines.components.transformers.encoders.one_hot.OneHotEncoder</code> method), 710	<code>save()</code> (<code>evalml.pipelines.components.transformers.PolynomialDetrender</code> method), 890
<code>save()</code> (<code>evalml.pipelines.components.transformers.encoders.one_hot.OneHotEncoder</code> method), 717	<code>save()</code> (<code>evalml.pipelines.components.transformers.preprocessing.datetime.DateTransformer</code> method), 762
<code>save()</code> (<code>evalml.pipelines.components.transformers.encoders.target.TargetEncoder</code> method), 714	<code>save()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DateTransformer</code> method), 793
<code>save()</code> (<code>evalml.pipelines.components.transformers.encoders.target.TargetEncoder</code> method), 720	<code>save()</code> (<code>evalml.pipelines.components.transformers.preprocessing.delayed.DelayedFeatureTransformer</code> method), 765
<code>save()</code> (<code>evalml.pipelines.components.transformers.features_selection.FeatureSelector</code> method), 723	<code>save()</code> (<code>evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer</code> method), 795

1453

- 1171
- save () (*evalml.pipelines.KNeighborsClassifier method*), 1174
- save () (*evalml.pipelines.LightGBMClassifier method*), 1177
- save () (*evalml.pipelines.LightGBMRegressor method*), 1180
- save () (*evalml.pipelines.LinearRegressor method*), 1182
- save () (*evalml.pipelines.LogisticRegressionClassifier method*), 1185
- save () (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method*), 1083
- save () (*evalml.pipelines.MulticlassClassificationPipeline method*), 1189
- save () (*evalml.pipelines.OneHotEncoder method*), 1192
- save () (*evalml.pipelines.PerColumnImputer method*), 1194
- save () (*evalml.pipelines.pipeline_base.PipelineBase method*), 1088
- save () (*evalml.pipelines.PipelineBase method*), 1198
- save () (*evalml.pipelines.ProphetRegressor method*), 1201
- save () (*evalml.pipelines.RandomForestClassifier method*), 1204
- save () (*evalml.pipelines.RandomForestRegressor method*), 1206
- save () (*evalml.pipelines.regression_pipeline.RegressionPipeline method*), 1093
- save () (*evalml.pipelines.RegressionPipeline method*), 1210
- save () (*evalml.pipelines.RFClassifierSelectFromModel method*), 1213
- save () (*evalml.pipelines.RFRegressorSelectFromModel method*), 1215
- save () (*evalml.pipelines.SimpleImputer method*), 1217
- save () (*evalml.pipelines.SklearnStackedEnsembleClassifier method*), 1220
- save () (*evalml.pipelines.SklearnStackedEnsembleRegressor method*), 1223
- save () (*evalml.pipelines.StandardScaler method*), 1225
- save () (*evalml.pipelines.SVMClassifier method*), 1227
- save () (*evalml.pipelines.SVMRegressor method*), 1230
- save () (*evalml.pipelines.TargetEncoder method*), 1232
- save () (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method*), 1100
- save () (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method*), 1105
- save () (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method*), 1111
- save () (*evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase method*), 1116
- save () (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method*), 1121
- save () (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline method*), 1237
- save () (*evalml.pipelines.TimeSeriesClassificationPipeline method*), 1243
- save () (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method*), 1248
- save () (*evalml.pipelines.TimeSeriesRegressionPipeline method*), 1253
- save () (*evalml.pipelines.Transformer method*), 1256
- save () (*evalml.pipelines.XGBoostClassifier method*), 1258
- save () (*evalml.pipelines.XGBoostRegressor method*), 1261
- save_plot () (in module *evalml.utils*), 1298
- save_plot () (in module *evalml.utils.gen_utils*), 1292
- scikit_learn_wrapped_estimator () (in module *evalml.pipelines.components.utils*), 922
- score () (*evalml.objectives.AccuracyBinary method*), 394
- score () (*evalml.objectives.AccuracyMulticlass method*), 396
- score () (*evalml.objectives.AUC method*), 398
- score () (*evalml.objectives.AUCMacro method*), 399
- score () (*evalml.objectives.AUCMicro method*), 401
- score () (*evalml.objectives.AUCWeighted method*), 403
- score () (*evalml.objectives.BalancedAccuracyBinary method*), 405
- score () (*evalml.objectives.BalancedAccuracyMulticlass method*), 406
- score () (*evalml.objectives.binary_classification_objective.BinaryClassificationObjective method*), 305
- score () (*evalml.objectives.BinaryClassificationObjective method*), 409
- score () (*evalml.objectives.cost_benefit_matrix.CostBenefitMatrix method*), 308
- score () (*evalml.objectives.CostBenefitMatrix method*), 412
- score () (*evalml.objectives.ExpVariance method*), 413
- score () (*evalml.objectives.F1 method*), 415
- score () (*evalml.objectives.F1Macro method*), 417
- score () (*evalml.objectives.F1Micro method*), 419
- score () (*evalml.objectives.F1Weighted method*), 420
- score () (*evalml.objectives.fraud_cost.FraudCost method*), 311
- score () (*evalml.objectives.fraud_cost.FraudCost method*), 423
- score () (*evalml.objectives.Gini method*), 426
- score () (*evalml.objectives.LeadScoring method*), 314
- score () (*evalml.objectives.LeadScoring method*), 428
- score () (*evalml.objectives.LogLossBinary method*), 420
- score () (*evalml.objectives.LogLossMulticlass method*), 420

[method](#)), 432
[score\(\)](#) ([evalml.objectives.MAE method](#)), 434
[score\(\)](#) ([evalml.objectives.MAPE method](#)), 435
[score\(\)](#) ([evalml.objectives.MaxError method](#)), 437
[score\(\)](#) ([evalml.objectives.MCCBinary method](#)), 439
[score\(\)](#) ([evalml.objectives.MCCMulticlass method](#)), 441
[score\(\)](#) ([evalml.objectives.MeanSquaredLogError method](#)), 443
[score\(\)](#) ([evalml.objectives.MedianAE method](#)), 444
[score\(\)](#) ([evalml.objectives.MSE method](#)), 446
[score\(\)](#) ([evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective method](#)), 316
[score\(\)](#) ([evalml.objectives.MulticlassClassificationObjective method](#)), 448
[score\(\)](#) ([evalml.objectives.objective_base.ObjectiveBase method](#)), 318
[score\(\)](#) ([evalml.objectives.ObjectiveBase method](#)), 450
[score\(\)](#) ([evalml.objectives.Precision method](#)), 452
[score\(\)](#) ([evalml.objectives.PrecisionMacro method](#)), 454
[score\(\)](#) ([evalml.objectives.PrecisionMicro method](#)), 456
[score\(\)](#) ([evalml.objectives.PrecisionWeighted method](#)), 457
[score\(\)](#) ([evalml.objectives.R2 method](#)), 459
[score\(\)](#) ([evalml.objectives.Recall method](#)), 461
[score\(\)](#) ([evalml.objectives.RecallMacro method](#)), 463
[score\(\)](#) ([evalml.objectives.RecallMicro method](#)), 464
[score\(\)](#) ([evalml.objectives.RecallWeighted method](#)), 466
[score\(\)](#) ([evalml.objectives.regression_objective.RegressionObjective method](#)), 321
[score\(\)](#) ([evalml.objectives.RegressionObjective method](#)), 468
[score\(\)](#) ([evalml.objectives.RootMeanSquaredError method](#)), 470
[score\(\)](#) ([evalml.objectives.RootMeanSquaredLogError method](#)), 471
[score\(\)](#) ([evalml.objectives.sensitivity_low_alert.SensitivityLowAlert method](#)), 323
[score\(\)](#) ([evalml.objectives.SensitivityLowAlert method](#)), 473
[score\(\)](#) ([evalml.objectives.standard_metrics.AccuracyBinary method](#)), 327
[score\(\)](#) ([evalml.objectives.standard_metrics.AccuracyMulticlass method](#)), 328
[score\(\)](#) ([evalml.objectives.standard_metrics.AUC method](#)), 331
[score\(\)](#) ([evalml.objectives.standard_metrics.AUCMacro method](#)), 332
[score\(\)](#) ([evalml.objectives.standard_metrics.AUCMicro method](#)), 334
[score\(\)](#) ([evalml.objectives.standard_metrics.AUCWeighted method](#)), 335
[score\(\)](#) ([evalml.objectives.standard_metrics.BalancedAccuracyBinary method](#)), 338
[score\(\)](#) ([evalml.objectives.standard_metrics.BalancedAccuracyMulticlass method](#)), 339
[score\(\)](#) ([evalml.objectives.standard_metrics.ExpVariance method](#)), 341
[score\(\)](#) ([evalml.objectives.standard_metrics.F1 method](#)), 343
[score\(\)](#) ([evalml.objectives.standard_metrics.F1Macro method](#)), 345
[score\(\)](#) ([evalml.objectives.standard_metrics.F1Micro method](#)), 346
[score\(\)](#) ([evalml.objectives.standard_metrics.F1Weighted method](#)), 348
[score\(\)](#) ([evalml.objectives.standard_metrics.Gini method](#)), 350
[score\(\)](#) ([evalml.objectives.standard_metrics.LogLossBinary method](#)), 352
[score\(\)](#) ([evalml.objectives.standard_metrics.LogLossMulticlass method](#)), 354
[score\(\)](#) ([evalml.objectives.standard_metrics.MAE method](#)), 356
[score\(\)](#) ([evalml.objectives.standard_metrics.MAPE method](#)), 357
[score\(\)](#) ([evalml.objectives.standard_metrics.MaxError method](#)), 359
[score\(\)](#) ([evalml.objectives.standard_metrics.MCCBinary method](#)), 361
[score\(\)](#) ([evalml.objectives.standard_metrics.MCCMulticlass method](#)), 363
[score\(\)](#) ([evalml.objectives.standard_metrics.MeanSquaredLogError method](#)), 364
[score\(\)](#) ([evalml.objectives.standard_metrics.MedianAE method](#)), 366
[score\(\)](#) ([evalml.objectives.standard_metrics.MSE method](#)), 367
[score\(\)](#) ([evalml.objectives.standard_metrics.Precision method](#)), 370
[score\(\)](#) ([evalml.objectives.standard_metrics.PrecisionMacro method](#)), 371
[score\(\)](#) ([evalml.objectives.standard_metrics.PrecisionMicro method](#)), 373
[score\(\)](#) ([evalml.objectives.standard_metrics.PrecisionWeighted method](#)), 374
[score\(\)](#) ([evalml.objectives.standard_metrics.R2 method](#)), 376
[score\(\)](#) ([evalml.objectives.standard_metrics.Recall method](#)), 378
[score\(\)](#) ([evalml.objectives.standard_metrics.RecallMacro method](#)), 380
[score\(\)](#) ([evalml.objectives.standard_metrics.RecallMicro method](#)), 382

`score()` (`evalml.objectives.standard_metrics.RecallWeighted` `property`), 409
`method`), 383 `score_needs_proba()`
`score()` (`evalml.objectives.standard_metrics.RootMeanSquaredError` `property`), 316
`method`), 385
`score()` (`evalml.objectives.standard_metrics.RootMeanSquaredLogError` `property`), 316
`method`), 386 `score_needs_proba()` (`evalml.objectives.MulticlassClassificationObjective`
`method`), 389 `score_needs_proba()`
`score()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` `property`), 319
`method`), 1068
`score()` (`evalml.pipelines.BinaryClassificationPipeline` `score_needs_proba()`
`method`), 1132 (`evalml.objectives.ObjectiveBase` `property`),
`score()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline`
`method`), 1075 `score_needs_proba()`
`score()` (`evalml.pipelines.ClassificationPipeline` (`evalml.objectives.regression_objective.RegressionObjective`
`method`), 1142 `property`), 321
`score()` (`evalml.pipelines.components.utils.WrappedSKClassifier` `score_needs_proba()`
`method`), 923 (`evalml.objectives.RegressionObjective` `prop-`
`score()` (`evalml.pipelines.components.utils.WrappedSKRegressor` `property`), 468
`method`), 924 `score_needs_proba()`
`score()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline`
`method`), 1083 `score_needs_proba()`
`score()` (`evalml.pipelines.MulticlassClassificationPipeline` `score_pipeline()` (in module
`method`), 1189 `evalml.automl.engine.engine_base`), 183
`score()` (`evalml.pipelines.pipeline_base.PipelineBase` `score_pipelines()`
`method`), 1088 (`evalml.automl.automl_search.AutoMLSearch`
`score()` (`evalml.pipelines.PipelineBase` `method`), 1199 `method`), 196
`score()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` `score_pipelines()` (`evalml.automl.AutoMLSearch`
`method`), 1093 `method`), 207
`score()` (`evalml.pipelines.RegressionPipeline` `method`), `score_pipelines()` (`evalml.AutoMLSearch`
1210 `method`), 1303
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesBinaryClassificationPipeline` `score_pipelines()`
`method`), 1100 `method`), 197
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` `score_pipelines()` (`evalml.AutoMLSearch`
`method`), 1106 `method`), 208
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline` `search()` (`evalml.AutoMLSearch` `method`), 1304
`method`), 1111 `search()` (in module `evalml.automl`), 209
`score()` (`evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase` `search()` (in module `evalml.automl.automl_search`),
`method`), 1116 197
`score()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` `search_iterative()` (in module `evalml`), 1305
`method`), 1122 (`evalml.automl.pipeline_search_plots.PipelineSearchPlots`
`score()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` `method`), 200
`method`), 1237 `search_iterative()` (in module `evalml.automl`),
`score()` (`evalml.pipelines.TimeSeriesClassificationPipeline` `search_iterative()` (in module `evalml.automl`),
`method`), 1243 210
`score()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` `search_iterative()` (in module
`method`), 1249 `evalml.automl.automl_search`), 198
`score()` (`evalml.pipelines.TimeSeriesRegressionPipeline` `SearchIterationPlot` (class in
`method`), 1253 `evalml.automl.pipeline_search_plots`), 200
`score_needs_proba()` `SEED_BOUNDS` (in module `evalml.utils`), 1298
(`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` `evalml.utils.gen_utils`), 1293
`property`), 305 `SelectByType` (class in
`score_needs_proba()` `evalml.pipelines.components`), 1025
(`evalml.objectives.BinaryClassificationObjective` `SelectByType` (class in

`evalml.pipelines.components.transformers), 896`

`SelectByType` (class in `evalml.pipelines.components.transformers.columns_selector`), 840

`SelectColumns` (class in `evalml.pipelines.components`), 1027

`SelectColumns` (class in `evalml.pipelines.components.transformers`), 898

`SelectColumns` (class in `evalml.pipelines.components.transformers.columns_selector`), 842

`send_data_to_cluster()` (`evalml.automl.engine.dask_engine.DaskEngine` method), 180

`send_data_to_cluster()` (`evalml.automl.engine.DaskEngine` method), 188

`SensitivityLowAlert` (class in `evalml.objectives`), 472

`SensitivityLowAlert` (class in `evalml.objectives.sensitivity_low_alert`), 322

`SequentialComputation` (class in `evalml.automl.engine.sequential_engine`), 185

`SequentialEngine` (class in `evalml.automl`), 211

`SequentialEngine` (class in `evalml.automl.engine`), 190

`SequentialEngine` (class in `evalml.automl.engine.sequential_engine`), 185

`set_fit()` (`evalml.pipelines.components.component_base_meta.ComponentBaseMeta` class method), 920

`set_fit()` (`evalml.pipelines.components.ComponentBaseMeta` class method), 941

`set_fit()` (`evalml.pipelines.components.transformers.ensemble_encoder.OracleEncoderMeta` class method), 711

`set_fit()` (`evalml.pipelines.components.transformers.imputer_target_imputer.TargetImputerMeta` class method), 749

`set_fit()` (`evalml.pipelines.pipeline_meta.PipelineBaseMeta` class method), 1089

`set_fit()` (`evalml.utils.base_meta.BaseMeta` class method), 1288

`set_params()` (`evalml.pipelines.components.utils.WrappedSKClassifier` method), 923

`set_params()` (`evalml.pipelines.components.utils.WrappedSKRegressor` method), 925

`setup_job_log()` (`evalml.automl.engine.cf_engine.CFEngine` static method), 178

`setup_job_log()` (`evalml.automl.engine.CFEngine` static method), 186

`setup_job_log()` (`evalml.automl.engine.dask_engine.DaskEngine` static method), 180

`setup_job_log()` (`evalml.automl.engine.DaskEngine` static method), 188

`setup_job_log()` (`evalml.automl.engine.engine_base.EngineBase` static method), 182

`setup_job_log()` (`evalml.automl.engine.EngineBase` static method), 189

`setup_job_log()` (`evalml.automl.engine.sequential_engine.SequentialEngine` static method), 185

`setup_job_log()` (`evalml.automl.engine.SequentialEngine` static method), 190

`setup_job_log()` (`evalml.automl.EngineBase` static method), 208

`setup_job_log()` (`evalml.automl.SequentialEngine` static method), 211

`silent_error_callback()` (in module `evalml.automl.callbacks`), 199

`SimpleImputer` (class in `evalml.pipelines`), 1216

`SimpleImputer` (class in `evalml.pipelines.components`), 1028

`SimpleImputer` (class in `evalml.pipelines.components.transformers`), 900

`SimpleImputer` (class in `evalml.pipelines.components.transformers.imputers`), 755

`SimpleImputer` (class in `evalml.pipelines.components.transformers.imputers.simple_imputer`), 744

`SklearnStackedEnsembleBase` (class in `evalml.pipelines.components.ensemble`), 483

`SklearnStackedEnsembleBase` (class in `evalml.pipelines.components.ensemble.sklearn_stacked_ensemble`), 474

`SklearnStackedEnsembleClassifier` (class in `evalml.pipelines`), 1218

`SklearnStackedEnsembleClassifier` (class in `evalml.pipelines.components`), 1031

`SklearnStackedEnsembleClassifier` (class in `evalml.pipelines.components.ensemble`), 485

`SklearnStackedEnsembleClassifier` (class in `evalml.pipelines.components.ensemble.sklearn_stacked_ensemble`), 477

`SklearnStackedEnsembleRegressor` (class in `evalml.pipelines`), 1220

`SklearnStackedEnsembleRegressor` (class in `evalml.pipelines.components`), 1033

`SklearnStackedEnsembleRegressor` (class in `evalml.pipelines.components.ensemble`), 488

`SklearnStackedEnsembleRegressor` (class in `evalml.pipelines.components.ensemble.sklearn_stacked_ensemble`), 480

`StackedEnsemble` (class in `evalml.tuners`), 1286

SKOPTuner (class in evalml.tuners.skopt_tuner),	1281	method),	182
sparsity_score() (evalml.data_checks.sparsity_data_check.SparsityDataCheck	static method), 233	submit_evaluation_job()	(evalml.automl.engine.EngineBase method),
sparsity_score() (evalml.data_checks.SparsityDataCheck	static method), 256	submit_evaluation_job()	189
SparsityDataCheck (class in evalml.data_checks),	256	(evalml.automl.engine.sequential_engine.SequentialEngine	method), 185
SparsityDataCheck (class in submit_evaluation_job()	evalml.data_checks.sparsity_data_check),	(evalml.automl.engine.SequentialEngine	method), 190
split() (evalml.preprocessing.data_splitters.time_series_split.TimeSeriesSplit	method), 1264	submit_evaluation_job()	(evalml.automl.EngineBase method), 208
split() (evalml.preprocessing.data_splitters.TimeSeriesSplit	method), 1267	submit_evaluation_job()	(evalml.automl.SequentialEngine method),
split() (evalml.preprocessing.data_splitters.training_validation_split.TrainingValidationSplit	method), 1265	submit_scoring_job()	
split() (evalml.preprocessing.data_splitters.TrainingValidationSplit	method), 1268	(evalml.automl.engine.cf_engine.CFEngine	method), 178
split() (evalml.preprocessing.TimeSeriesSplit	method), 1272	submit_scoring_job()	(evalml.automl.engine.CFEngine method),
split() (evalml.preprocessing.TrainingValidationSplit	method), 1272	submit_scoring_job()	187
split_data() (in module evalml.preprocessing),	1271	(evalml.automl.engine.dask_engine.DaskEngine	method), 180
split_data() (in module evalml.preprocessing.utils),	1269	submit_scoring_job()	(evalml.automl.engine.DaskEngine method),
StandardScaler (class in evalml.pipelines),	1223	188	
StandardScaler (class in submit_scoring_job()	evalml.pipelines.components), 1036	(evalml.automl.engine.engine_base.EngineBase	method), 182
StandardScaler (class in submit_scoring_job()	evalml.pipelines.components.transformers),	902	submit_scoring_job()
StandardScaler (class in submit_scoring_job()	evalml.pipelines.components.transformers.scalers),	834	submit_scoring_job()
StandardScaler (class in submit_scoring_job()	evalml.pipelines.components.transformers.scalers.standard_scaler),	832	submit_scoring_job()
submit() (evalml.automl.engine.cf_engine.CFClient	method), 176	(evalml.automl.engine.sequential_engine.SequentialEngine	method), 190
submit_evaluation_job()	(evalml.automl.engine.cf_engine.CFEngine	method), 178	submit_scoring_job()
submit_evaluation_job()	(evalml.automl.engine.CFEngine method),	186	submit_training_job()
submit_evaluation_job()	(evalml.automl.engine.dask_engine.DaskEngine	method), 180	submit_training_job()
submit_evaluation_job()	(evalml.automl.engine.DaskEngine method),	188	submit_training_job()
submit_evaluation_job()	(evalml.automl.engine.engine_base.EngineBase	submit_training_job()	

TargetDistributionDataCheck	(class in evalml.data_checks), 257	evalml.pipelines.components.transformers.preprocessing.text_transformer , 784
TargetDistributionDataCheck	(class in evalml.data_checks.target_distribution_data_check), 235	threshold() (evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline property), 1069
TargetEncoder	(class in evalml.pipelines), 1230	threshold() (evalml.pipelines.binary_classification_pipeline_mixin.BinaryClassificationPipelineMixin property), 1070
TargetEncoder	(class in evalml.pipelines.components), 1043	threshold() (evalml.pipelines.BinaryClassificationPipeline property), 1133
TargetEncoder	(class in evalml.pipelines.components.transformers), 904	threshold() (evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline property), 1100
TargetEncoder	(class in evalml.pipelines.components.transformers.encoders), 718	threshold() (evalml.pipelines.TimeSeriesBinaryClassificationPipeline property), 1238
TargetEncoder	(class in evalml.pipelines.components.transformers.encoders.target_encoder), 712	time_elapsed() (in module evalml.utils.logger), 1293
TargetImputer	(class in evalml.pipelines.components), 1045	TimeSeriesBaselineEstimator (class in evalml.pipelines.components), 1050
TargetImputer	(class in evalml.pipelines.components.transformers), 906	TimeSeriesBaselineEstimator (class in evalml.pipelines.components.estimators), 689
TargetImputer	(class in evalml.pipelines.components.transformers.imputers), 757	TimeSeriesBaselineEstimator (class in evalml.pipelines.components.estimators.regressors), 623
TargetImputer	(class in evalml.pipelines.components.transformers.imputers), 747	TimeSeriesBaselineEstimator (class in evalml.pipelines.components.estimators.regressors.time_series_regression), 588
TargetImputerMeta	(class in evalml.pipelines.components.transformers.imputers), 749	TimeSeriesBinaryClassificationPipeline (class in evalml.pipelines), 1232
TargetLeakageDataCheck	(class in evalml.data_checks), 258	TimeSeriesBinaryClassificationPipeline (class in evalml.pipelines.time_series_classification_pipelines), 1094
TargetLeakageDataCheck	(class in evalml.data_checks.target_leakage_data_check), 236	TimeSeriesBinaryClassificationPipeline (class in evalml.pipelines), 1238
TargetTransformer	(class in evalml.pipelines.components.transformers.transformer), 845	TimeSeriesClassificationPipeline (class in evalml.pipelines.time_series_classification_pipelines), 1100
TextFeaturizer	(class in evalml.pipelines.components), 1047	TimeSeriesClassificationPipeline (class in evalml.pipelines.time_series_classification_pipelines), 1244
TextFeaturizer	(class in evalml.pipelines.components.transformers), 909	TimeSeriesMulticlassClassificationPipeline (class in evalml.pipelines), 1106
TextFeaturizer	(class in evalml.pipelines.components.transformers.preprocessing.text_featurizer), 811	TimeSeriesMulticlassClassificationPipeline (class in evalml.pipelines.time_series_classification_pipelines), 1112
TextFeaturizer	(class in evalml.pipelines.components.transformers.preprocessing.text_featurizer), 781	TimeSeriesRegressionObjective (class in evalml.objectives.time_series_regression_objective), 387
TextTransformer	(class in evalml.pipelines.components.transformers.preprocessing), 813	TimeSeriesRegressionPipeline (class in evalml.pipelines), 1249
TextTransformer	(class in evalml.pipelines.components.transformers.preprocessing), 813	TimeSeriesRegressionPipeline (class in evalml.pipelines.time_series_regression_pipeline), 1117
TextTransformer	(class in evalml.pipelines.components.transformers.preprocessing), 813	TimeSeriesSplit (class in evalml.preprocessing), 1271
TextTransformer	(class in evalml.pipelines.components.transformers.preprocessing), 813	TimeSeriesSplit (class in evalml.preprocessing.data_splitters), 1267

TimeSeriesSplit (class in method), 944
 evalml.preprocessing.data_splitters.time_series_split, 1264
 to_dict() (evalml.data_checks.data_check_action.DataCheckAction method), 215
 to_dict() (evalml.data_checks.data_check_message.DataCheckError method), 216
 to_dict() (evalml.data_checks.data_check_message.DataCheckMessage method), 217
 to_dict() (evalml.data_checks.data_check_message.DataCheckWarning method), 217
 to_dict() (evalml.data_checks.DataCheckAction method), 243
 to_dict() (evalml.data_checks.DataCheckError method), 244
 to_dict() (evalml.data_checks.DataCheckMessage method), 244
 to_dict() (evalml.data_checks.DataCheckWarning method), 246
 train_and_score_pipeline() (in module *evalml.automl.engine*), 190
 train_and_score_pipeline() (in module *evalml.automl.engine.engine_base*), 184
 train_pipeline() (in module *evalml.automl.engine*), 191
 train_pipeline() (in module *evalml.automl.engine.engine_base*), 184
 train_pipelines() (evalml.automl.automl_search.AutoMLSearch method), 197
 train_pipelines() (evalml.automl.AutoMLSearch method), 208
 train_pipelines() (evalml.AutoMLSearch method), 1304
 TrainingValidationSplit (class in method), 1272
 evalml.preprocessing.data_splitters, 1268
 TrainingValidationSplit (class in method), 1026
 evalml.preprocessing.data_splitters.training_validation_split, 1265
 transform() (evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline method), 1069
 transform() (evalml.pipelines.BinaryClassificationPipeline method), 1133
 transform() (evalml.pipelines.classification_pipeline.ClassificationPipeline method), 1075
 transform() (evalml.pipelines.ClassificationPipeline method), 1143
 transform() (evalml.pipelines.component_graph.ComponentGraph method), 1078
 transform() (evalml.pipelines.ComponentGraph method), 1145
 transform() (evalml.pipelines.components.DateTimeFeaturizer method), 944
 transform() (evalml.pipelines.components.DelayedFeatureTransformer method), 952
 transform() (evalml.pipelines.components.DFSTransformer method), 954
 transform() (evalml.pipelines.components.DropColumns method), 956
 transform() (evalml.pipelines.components.DropNullColumns method), 958
 transform() (evalml.pipelines.components.DropRowsTransformer method), 960
 transform() (evalml.pipelines.components.EmailFeaturizer method), 967
 transform() (evalml.pipelines.components.FeatureSelector method), 978
 transform() (evalml.pipelines.components.Imputer method), 980
 transform() (evalml.pipelines.components.LinearDiscriminantAnalysis method), 990
 transform() (evalml.pipelines.components.LogTransformer method), 998
 transform() (evalml.pipelines.components.LSA method), 1000
 transform() (evalml.pipelines.components.OneHotEncoder method), 1003
 transform() (evalml.pipelines.components.Oversampler method), 1005
 transform() (evalml.pipelines.components.PCA method), 1007
 transform() (evalml.pipelines.components.PerColumnImputer method), 1010
 transform() (evalml.pipelines.components.PolynomialDetrender method), 1012
 transform() (evalml.pipelines.components.RFClassifierSelectFromModel method), 1022
 transform() (evalml.pipelines.components.RFRegressorSelectFromModel method), 1024
 transform() (evalml.pipelines.components.SelectByType method), 1026
 transform() (evalml.pipelines.components.SelectColumns method), 1028
 transform() (evalml.pipelines.components.SimpleImputer method), 1031
 transform() (evalml.pipelines.components.StandardScaler method), 1038
 transform() (evalml.pipelines.components.TargetEncoder method), 1045
 transform() (evalml.pipelines.components.TargetImputer method), 1047
 transform() (evalml.pipelines.components.TextFeaturizer method), 1049
 transform() (evalml.pipelines.components.Transformer method), 1054
 transform() (evalml.pipelines.components.transformers.column_selector.ColumnSelector method), 1056

`method`), 838
`transform()` (`evalml.pipelines.components.transformers.column_selectors.DateFeatureTransformer`), 840
`method`), 840
`transform()` (`evalml.pipelines.components.transformers.column_selectors.ScaledBinaryType`), 842
`method`), 842
`transform()` (`evalml.pipelines.components.transformers.column_selectors.ScaledCategorical`), 844
`method`), 844
`transform()` (`evalml.pipelines.components.transformers.column_selectors.SimpleDateTimeFeatureTransformer`), 853
`method`), 853
`transform()` (`evalml.pipelines.components.transformers.column_selectors.DelayedFeatureTransformer`), 855
`method`), 855
`transform()` (`evalml.pipelines.components.transformers.target_encoders.DESTransformer`), 858
`method`), 858
`transform()` (`evalml.pipelines.components.transformers.target_encoders.TargetEncoderWithInteractionEffects`), 699
`method`), 699
`transform()` (`evalml.pipelines.components.transformers.target_encoders.LinearDiscriminantAnalysis`), 704
`method`), 704
`transform()` (`evalml.pipelines.components.transformers.target_encoders.LogTransformedLinearDiscriminantAnalysis`), 707
`method`), 707
`transform()` (`evalml.pipelines.components.transformers.target_encoders.LSA`), 702
`method`), 702
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.OneHotEncoder`), 860
`method`), 860
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.Oversampler`), 862
`method`), 862
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.PCA`), 864
`method`), 864
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.PerColumnImputer`), 866
`method`), 866
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.PolynomialDet`), 710
`method`), 710
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.Preprocessing`), 717
`method`), 717
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.TargetEncoder`), 714
`method`), 714
`transform()` (`evalml.pipelines.components.transformers.one_hot_encoders.Preprocessing`), 720
`method`), 720
`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 723
`method`), 723
`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 732
`method`), 732
`transform()` (`evalml.pipelines.components.transformers.feature_selection.RFClassifierPreprocessor`), 726
`method`), 726
`transform()` (`evalml.pipelines.components.transformers.feature_selection.RFRegressorPreprocessor`), 729
`method`), 729
`transform()` (`evalml.pipelines.components.transformers.feature_selection.RFClassifierPreprocessor`), 735
`method`), 735
`transform()` (`evalml.pipelines.components.transformers.feature_selection.RFRegressorPreprocessor`), 737
`method`), 737
`transform()` (`evalml.pipelines.components.transformers.feature_selection.RFClassifierPreprocessor`), 869
`method`), 869
`transform()` (`evalml.pipelines.components.transformers.imputers.Imputer`), 871
`method`), 871
`transform()` (`evalml.pipelines.components.transformers.imputers.Imputer`), 871
`method`), 871

1463

[method](#)), 1112
[transform\(\)](#) ([evalml.pipelines.time_series_pipeline_base.TimeSeriesPipelineBase](#)
[method](#)), 1117
[transform\(\)](#) ([evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline](#)), 1057
[method](#)), 1122
[transform\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#)), 915
[method](#)), 1238
[transform\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#)), 915
[method](#)), 1243
[transform\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#)), 915
[method](#)), 1249
[transform\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#)), 788
[method](#)), 1254
[transform\(\)](#) ([evalml.pipelines.Transformer](#) [method](#)), 1256
[Transformer](#) ([class](#) in [evalml.pipelines](#)), 1254
[Transformer](#) ([class](#) in [evalml.pipelines.components](#)), 1052
[Transformer](#) ([class](#) in [evalml.pipelines.components.transformers](#)), 911
[Transformer](#) ([class](#) in [evalml.pipelines.components.transformers.transformer](#)), 847
[tune_binary_threshold\(\)](#) ([in](#) [module](#) [evalml.automl](#)), 211
[tune_binary_threshold\(\)](#) ([in](#) [module](#) [evalml.automl.utils](#)), 202
[Tuner](#) ([class](#) in [evalml.tuners](#)), 1286
[Tuner](#) ([class](#) in [evalml.tuners.tuner](#)), 1282

U

[Undersampler](#) ([class](#) in [evalml.pipelines.components](#)), 1054
[Undersampler](#) ([class](#) in [evalml.pipelines.components.transformers](#)), 913
[Undersampler](#) ([class](#) in [evalml.pipelines.components.transformers.samplers](#)), 828
[Undersampler](#) ([class](#) in [evalml.pipelines.components.transformers.samplers.undersampler](#)), 823
[uniqueness_score\(\)](#) ([evalml.data_checks.uniqueness_data_check.UniquenessDataCheck](#)
[static method](#)), 238
[uniqueness_score\(\)](#) ([evalml.data_checks.UniquenessDataCheck](#)
[static method](#)), 259
[UniquenessDataCheck](#) ([class](#) in [evalml.data_checks](#)), 259
[UniquenessDataCheck](#) ([class](#) in [evalml.data_checks.uniqueness_data_check](#)), 237

[update\(\)](#) ([evalml.automl.pipeline_search_plots.SearchIterationPlot](#)
[method](#)), 208
[URLFeaturizer](#) ([class](#) in [evalml.pipelines.components.transformers.preprocessing](#)), 1057
[URLFeaturizer](#) ([class](#) in [evalml.pipelines.components.transformers.preprocessing.transformers](#)), 788
[URLFeaturizer](#) ([class](#) in [evalml.pipelines.components.transformers.preprocessing.transformers.url_featurizer](#)), 788

V

[validate\(\)](#) ([evalml.data_checks.class_imbalance_data_check.ClassImbalanceDataCheck](#)
[method](#)), 212
[validate\(\)](#) ([evalml.data_checks.ClassImbalanceDataCheck](#)
[method](#)), 241
[validate\(\)](#) ([evalml.data_checks.data_check.DataCheck](#)
[method](#)), 214
[validate\(\)](#) ([evalml.data_checks.data_checks.DataChecks](#)
[method](#)), 220
[validate\(\)](#) ([evalml.data_checks.DataCheck](#) [method](#)), 242
[validate\(\)](#) ([evalml.data_checks.DataChecks](#)
[method](#)), 246
[validate\(\)](#) ([evalml.data_checks.datetime_format_data_check.DateTimeFormatDataCheck](#)
[method](#)), 221
[validate\(\)](#) ([evalml.data_checks.datetime_nan_data_check.DateTimeNaNDataCheck](#)
[method](#)), 222
[validate\(\)](#) ([evalml.data_checks.DateTimeFormatDataCheck](#)
[method](#)), 247
[validate\(\)](#) ([evalml.data_checks.DateTimeNaNDataCheck](#)
[method](#)), 247
[validate\(\)](#) ([evalml.data_checks.default_data_checks.DefaultDataChecks](#)
[method](#)), 224
[validate\(\)](#) ([evalml.data_checks.DefaultDataChecks](#)
[method](#)), 249
[validate\(\)](#) ([evalml.data_checks.EmptyDataChecks](#)
[method](#)), 249
[validate\(\)](#) ([evalml.data_checks.highly_null_data_check.HighlyNullDataCheck](#)
[method](#)), 225
[validate\(\)](#) ([evalml.data_checks.HighlyNullDataCheck](#)
[method](#)), 250
[validate\(\)](#) ([evalml.data_checks.id_columns_data_check.IDColumnsDataCheck](#)
[method](#)), 226
[validate\(\)](#) ([evalml.data_checks.IDColumnsDataCheck](#)
[method](#)), 251
[validate\(\)](#) ([evalml.data_checks.invalid_targets_data_check.InvalidTargetsDataCheck](#)
[method](#)), 228
[validate\(\)](#) ([evalml.data_checks.InvalidTargetDataCheck](#)
[method](#)), 252

`method`), 443
`validate_inputs()` (`evalml.objectives.MedianAE` `method`), 444
`validate_inputs()` (`evalml.objectives.MSE` `method`), 446
`validate_inputs()` (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` `method`), 316
`validate_inputs()` (`evalml.objectives.MulticlassClassificationObjective` `method`), 448
`validate_inputs()` (`evalml.objectives.objective_base.ObjectiveBase` `method`), 319
`validate_inputs()` (`evalml.objectives.ObjectiveBase` `method`), 450
`validate_inputs()` (`evalml.objectives.Precision` `method`), 452
`validate_inputs()` (`evalml.objectives.PrecisionMacro` `method`), 454
`validate_inputs()` (`evalml.objectives.PrecisionMicro` `method`), 456
`validate_inputs()` (`evalml.objectives.PrecisionWeighted` `method`), 457
`validate_inputs()` (`evalml.objectives.R2` `method`), 459
`validate_inputs()` (`evalml.objectives.Recall` `method`), 461
`validate_inputs()` (`evalml.objectives.RecallMacro` `method`), 463
`validate_inputs()` (`evalml.objectives.RecallMicro` `method`), 464
`validate_inputs()` (`evalml.objectives.RecallWeighted` `method`), 466
`validate_inputs()` (`evalml.objectives.regression_objective.RegressionObjective` `method`), 321
`validate_inputs()` (`evalml.objectives.RegressionObjective` `method`), 468
`validate_inputs()` (`evalml.objectives.RootMeanSquaredError` `method`), 470
`validate_inputs()` (`evalml.objectives.RootMeanSquaredLogError` `method`), 471
`validate_inputs()` (`evalml.objectives.sensitivity_low_alert.SensitivityLowAlert` `method`), 324
`validate_inputs()` (`evalml.objectives.SensitivityLowAlert` `method`), 473
`validate_inputs()` (`evalml.objectives.standard_metrics.AccuracyBinary` `method`), 327
`validate_inputs()` (`evalml.objectives.standard_metrics.AccuracyMulticlass` `method`), 329
`validate_inputs()` (`evalml.objectives.standard_metrics.AUC` `method`), 331
`validate_inputs()` (`evalml.objectives.standard_metrics.AUCMacro` `method`), 332
`validate_inputs()` (`evalml.objectives.standard_metrics.AUCMicro` `method`), 334
`validate_inputs()` (`evalml.objectives.standard_metrics.AUCWeighted` `method`), 336
`validate_inputs()` (`evalml.objectives.standard_metrics.BalancedAccuracyBinary` `method`), 338
`validate_inputs()` (`evalml.objectives.standard_metrics.BalancedAccuracyMulticlass` `method`), 339
`validate_inputs()` (`evalml.objectives.standard_metrics.ExpVariance` `method`), 341
`validate_inputs()` (`evalml.objectives.standard_metrics.F1` `method`), 343
`validate_inputs()` (`evalml.objectives.standard_metrics.F1Macro` `method`), 345
`validate_inputs()` (`evalml.objectives.standard_metrics.F1Micro` `method`), 346
`validate_inputs()` (`evalml.objectives.standard_metrics.F1Weighted` `method`), 348
`validate_inputs()` (`evalml.objectives.standard_metrics.Gini` `method`), 350
`validate_inputs()` (`evalml.objectives.standard_metrics.LogLossBinary` `method`), 352
`validate_inputs()` (`evalml.objectives.standard_metrics.LogLossMulticlass` `method`), 354
`validate_inputs()` (`evalml.objectives.standard_metrics.MAE` `method`), 356

[method](#)), 356
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MAPE](#) [method](#)), 357
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MaxError](#) [method](#)), 359
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MCCBinary](#) [method](#)), 361
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MCCMulticlass](#) [method](#)), 363
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MeanSquaredLogError](#) [method](#)), 364
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MedianAE](#) [method](#)), 366
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.MSE](#) [method](#)), 368
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.Precision](#) [method](#)), 370
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.PrecisionMacro](#) [method](#)), 371
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.PrecisionMicro](#) [method](#)), 373
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.PrecisionWeighted](#) [method](#)), 375
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.R2](#) [method](#)), 376
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.Recall](#) [method](#)), 378
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.RecallMacro](#) [method](#)), 380
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.RecallMicro](#) [method](#)), 382
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.RecallWeighted](#) [method](#)), 383
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.RootMeanSquaredError](#) [method](#)), 385
[validate_inputs\(\)](#) ([evalml.objectives.standard_metrics.RootMeanSquaredLogError](#) [method](#)), 387
[validate_inputs\(\)](#) ([evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective](#) [method](#)), 389
[value\(\)](#) ([evalml.data_checks.data_check_action_code.DataCheckActionCode](#) [method](#)), 215
[value\(\)](#) ([evalml.data_checks.data_check_message_code.DataCheckMessageCode](#) [method](#)), 219
[value\(\)](#) ([evalml.data_checks.data_check_message_type.DataCheckMessageType](#) [method](#)), 219
[value\(\)](#) ([evalml.data_checks.DataCheckActionCode](#) [method](#)), 243
[value\(\)](#) ([evalml.data_checks.DataCheckMessageCode](#) [method](#)), 245
[value\(\)](#) ([evalml.data_checks.DataCheckMessageType](#) [method](#)), 246
[value\(\)](#) ([evalml.exceptions.exceptions.PartialDependenceErrorCode](#) [method](#)), 266
[value\(\)](#) ([evalml.exceptions.PartialDependenceErrorCode](#) [method](#)), 268
[value\(\)](#) ([evalml.model_family.model_family.ModelFamily](#) [method](#)), 270
[value\(\)](#) ([evalml.model_family.ModelFamily](#) [method](#)), 271
[value\(\)](#) ([evalml.model_understanding.prediction_explanations.explainer.Explainer](#) [method](#)), 275
[value\(\)](#) ([evalml.problem_types.problem_types.ProblemTypes](#) [method](#)), 1274
[value\(\)](#) ([evalml.problem_types.ProblemTypes](#) [method](#)), 1278
[visualize_decision_tree\(\)](#) (in module [evalml.model_understanding.graphs](#)), 288
W
[warning\(\)](#) ([evalml.automl.engine.engine_base.JobLogger](#) [method](#)), 183
[warning_not_unique_enough](#) (in module [evalml.data_checks.uniqueness_data_check](#)), 239
[warning_too_unique](#) (in module [evalml.data_checks.sparsity_data_check](#)), 234
[warning_too_unique](#) (in module [evalml.data_checks.uniqueness_data_check](#)), 239
[WrappedSKClassifier](#) (class in [evalml.pipelines.components.utils](#)), 922
[WrappedSKRegressor](#) (class in [evalml.pipelines.components.utils](#)), 923
[write_to_logger\(\)](#) ([evalml.automl.engine.engine_base.JobLogger](#) [method](#)), 183

X

XGBoostClassifier (class in *evalml.pipelines*),
[1256](#)

XGBoostClassifier (class in
evalml.pipelines.components), [1059](#)

XGBoostClassifier (class in
evalml.pipelines.components.estimators),
[692](#)

XGBoostClassifier (class in
evalml.pipelines.components.estimators.classifiers),
[552](#)

XGBoostClassifier (class in
evalml.pipelines.components.estimators.classifiers.xgboost_classifier),
[521](#)

XGBoostRegressor (class in *evalml.pipelines*), [1259](#)

XGBoostRegressor (class in
evalml.pipelines.components), [1061](#)

XGBoostRegressor (class in
evalml.pipelines.components.estimators),
[694](#)

XGBoostRegressor (class in
evalml.pipelines.components.estimators.regressors),
[625](#)

XGBoostRegressor (class in
evalml.pipelines.components.estimators.regressors.xgboost_regressor),
[591](#)