
EvalML Documentation

Release 0.29.0

Alteryx, Inc.

Jul 22, 2021

CONTENTS

1	Install	3
2	Start	5
3	Tutorials	13
4	User Guide	41
5	API Reference	133
6	Release Notes	1205
	Python Module Index	1247
	Index	1251

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 Pip with all dependencies

To install evalml with pip, run the following command:

```
pip install evalml
```

1.2 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
```

1.2.1 Add-ons

Update checker Receive automatic notifications of new EvalML releases

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

1.3 Conda with all dependencies

To install evalml with conda run the following command:

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

1.4 Conda with core dependencies

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

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

1.5 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.

1.6 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
```

1.7 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	category	Categorical	['category']
customer_present	bool	Boolean	[]
expiration_date	category	Categorical	['category']
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)
```

```
Generating pipelines to search over...
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: random_forest, linear_model, xgboost, decision_tree, catboost,
↪ extra_trees, 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/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler + Standard Scaler:

Starting cross validation
Finished cross validation - mean F1: 0.263
High coefficient of variation (cv >= 0.2) within cross validation scores.

Elastic Net Classifier w/ Imputer + DateTime Featurization Component + One
↪Hot Encoder + SMOTENC Oversampler + Standard Scaler may not perform as estimated on
↪unseen data.

Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler:

Starting cross validation
Finished cross validation - mean F1: 0.448
High coefficient of variation (cv >= 0.2) within cross validation scores.

Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One
↪Hot Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.

Random Forest Classifier w/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler:

Starting cross validation
Finished cross validation - mean F1: 0.500

LightGBM Classifier w/ Imputer + DateTime Featurization Component + One Hot Encoder +
↪SMOTENC Oversampler:

Starting cross validation
Finished cross validation - mean F1: 0.417
High coefficient of variation (cv >= 0.2) within cross validation scores.

LightGBM Classifier w/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.

Logistic Regression Classifier w/ Imputer + DateTime Featurization Component + One
↪Hot Encoder + SMOTENC Oversampler + Standard Scaler:

Starting cross validation
Finished cross validation - mean F1: 0.233

(continues on next page)

(continued from previous page)

```

[14:31:35] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
↪behavior.
[14:31:35] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
↪behavior.
[14:31:36] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
↪behavior.
XGBoost Classifier w/ Imputer + DateTime Featurization Component + One Hot Encoder +
↪SMOTENC Oversampler:
  Starting cross validation
  Finished cross validation - mean F1: 0.372
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  XGBoost Classifier w/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.
Extra Trees Classifier w/ Imputer + DateTime Featurization Component + One Hot
↪Encoder + SMOTENC Oversampler:
  Starting cross validation
  Finished cross validation - mean F1: 0.263
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Extra Trees Classifier w/ Imputer + DateTime Featurization Component + One
↪Hot Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.
CatBoost Classifier w/ Imputer + DateTime Featurization Component + SMOTENC
↪Oversampler:
  Starting cross validation
  Finished cross validation - mean F1: 0.646
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  CatBoost Classifier w/ Imputer + DateTime Featurization Component + SMOTENC
↪Oversampler may not perform as estimated on unseen data.

Search finished after 00:19
Best pipeline: CatBoost Classifier w/ Imputer + DateTime Featurization Component +
↪SMOTENC Oversampler
Best pipeline F1: 0.646154

```

We also provide a *standalone* ``search`` method [../generated/evalml.automl.search.html](https://evalml.alteryx.com/en/latest/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.

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

```

[7]:   id  pipeline_name  search_order  \
0   8  CatBoost Classifier w/ Imputer + DateTime Feat...      8
1   3  Random Forest Classifier w/ Imputer + DateTime...      3
2   2  Decision Tree Classifier w/ Imputer + DateTime...      2
3   4  LightGBM Classifier w/ Imputer + DateTime Feat...      4
4   6  XGBoost Classifier w/ Imputer + DateTime Featu...      6
5   7  Extra Trees Classifier w/ Imputer + DateTime F...      7

```

(continues on next page)

(continued from previous page)

```

6  1 Elastic Net Classifier w/ Imputer + DateTime F... 1
7  5 Logistic Regression Classifier w/ Imputer + Da... 5
8  0 Mode Baseline Binary Classification Pipeline 0

mean_cv_score standard_deviation_cv_score validation_score \
0 0.646154 0.213175 0.769231
1 0.500000 0.100000 0.400000
2 0.447992 0.328221 0.769231
3 0.416667 0.144338 0.500000
4 0.372222 0.256219 0.666667
5 0.263091 0.233224 0.444444
6 0.262500 0.073006 0.266667
7 0.232828 0.044955 0.266667
8 0.000000 0.000000 0.000000

percent_better_than_baseline high_variance_cv \
0 64.615385 True
1 50.000000 False
2 44.799226 True
3 41.666667 True
4 37.222222 True
5 26.309068 True
6 26.250000 True
7 23.282828 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'}}

```

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

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

```

*****
* Random Forest Classifier w/ Imputer + DateTime Featurization Component + One Hot_
↳Encoder + SMOTENC Oversampler *
*****

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Imputer
    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None

```

(continues on next page)

(continued from previous page)

```

    * numeric_fill_value : None
2. DateTime Featurization Component
    * features_to_extract : ['year', 'month', 'day_of_week', 'hour']
    * encode_as_categories : False
    * date_index : None
3. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
4. SMOTENC Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * categorical_features : [3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
↪22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
↪43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59]
    * k_neighbors : 5
5. Random Forest Classifier
    * n_estimators : 100
    * max_depth : 6
    * n_jobs : -1

```

Training

```
=====
```

Training for binary problems.

Objective to optimize binary classification pipeline thresholds for: <evalml.

↪objectives.standard_metrics.F1 object at 0x7f2397a18b80>

Total training time (including CV): 2.4 seconds

Cross Validation

```
-----
```

	F1	MCC	Binary	Log	Loss	Binary	AUC	Precision	Balanced	Accuracy
↪Binary	Accuracy	Binary	# Training	# Validation						
0	0.400	0.476		0.264	0.824		1.000		0.	
↪625		0.910	133	67						
1	0.500	0.490		0.333	0.712		0.750		0.	
↪679		0.910	133	67						
2	0.600	0.634		0.260	0.879		1.000		0.	
↪714		0.939	134	66						
mean	0.500	0.533		0.286	0.805		0.917		0.	
↪673		0.920	-	-						
std	0.100	0.087		0.041	0.085		0.144		0.	
↪045		0.017	-	-						
coef of var	0.200	0.163		0.144	0.106		0.157		0.	
↪067		0.018	-	-						

We can also view the pipeline parameters directly:

```

[9]: pipeline = automl.get_pipeline(3)
print(pipeline.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',
10'handle_unknown': 'ignore', 'handle_missing': 'error'}, 'SMOTENC Oversampler': {'sampling_ratio': 0.25, 'k_neighbors_default': 5, 'n_jobs': -1, 'sampling_ratio_dict':
↪None, 'categorical_features': [3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
↪22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
↪43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59], 'k_neighbors':

```

(continued from previous page)

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

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

[10]: OrderedDict([('F1', 0.5)])
```

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

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

[11]:
```


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]:
```

	Physical Type	Logical Type	Semantic Tag(s)
Column			
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)
```

```
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: linear_model, 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 Fraud Cost: 0.790
```

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

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

```
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.286
  High coefficient of variation (cv >= 0.2) within cross validation scores.
```

```
  Elastic Net Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +
↳Standard Scaler may not perform as estimated on unseen data.
```

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

```
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.289
  High coefficient of variation (cv >= 0.2) within cross validation scores.
```

```
  Random Forest Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler
↳may not perform as estimated on unseen data.
```

```
Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +
↳Standard Scaler:
```

```
  Starting cross validation
  Finished cross validation - mean Fraud Cost: 0.286
  High coefficient of variation (cv >= 0.2) within cross validation scores.
```

```
  Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC
↳Oversampler + Standard Scaler may not perform as estimated on unseen data.
```

```
Search finished after 00:13
```

```
Best pipeline: Elastic Net Classifier w/ Imputer + One Hot Encoder + SMOTENC
↳Oversampler + Standard Scaler
```

```
Best pipeline Fraud Cost: 0.285505
```

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    1  Elastic Net Classifier w/ Imputer + One Hot En...      1
1    3  Logistic Regression Classifier w/ Imputer + On...      3
2    2  Random Forest Classifier w/ Imputer + One Hot ...      2
3    0      Mode Baseline Binary Classification Pipeline      0

      mean_cv_score  standard_deviation_cv_score  validation_score  \
0          0.285505              0.436205          0.040953
1          0.285505              0.436205          0.040953
2          0.289310              0.434436          0.790953
3          0.789648              0.001136          0.790953

      percent_better_than_baseline  high_variance_cv  \
0                      50.414342              True
1                      50.414342              True
2                      50.033835              True
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"])
```

```
*****
* Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC 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
```

(continues on next page)

(continued from previous page)

```

    * 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. SMOTENC Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * categorical_features : [3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
    * k_neighbors : 5
4. Standard Scaler
5. Logistic Regression Classifier
    * penalty : l2
    * C : 1.0
    * n_jobs : -1
    * multi_class : auto
    * solver : lbfgs

```

Training

=====

Training for binary problems.

Objective to optimize binary classification pipeline thresholds for: <evalml.

→objectives.fraud_cost.FraudCost object at 0x7f9559214790>

Total training time (including CV): 3.5 seconds

Cross Validation

	Fraud Cost	AUC	F1	Precision	# Training	# Validation
0	0.041	0.854	0.011	1.000	2,666	1,334
1	0.789	0.799	0.000	0.000	2,667	1,333
2	0.026	0.828	0.355	0.228	2,667	1,333
mean	0.286	0.827	0.122	0.409	-	-
std	0.436	0.027	0.202	0.524	-	-
coef of var	1.528	0.033	1.656	1.280	-	-

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', 0.8097111537038906),
                  ('Fraud Cost', 0.0015780080694855455)])
```

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)

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: linear_model, 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 + SMOTENC Oversampler + Standard_
↪Scaler:
Starting cross validation
Finished cross validation - mean AUC: 0.827
Random Forest Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler:
Starting cross validation
Finished cross validation - mean AUC: 0.853
Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +_
↪Standard Scaler:
Starting cross validation
Finished cross validation - mean AUC: 0.827

Search finished after 00:07
Best pipeline: Random Forest Classifier w/ Imputer + One Hot Encoder + SMOTENC_
↪Oversampler

(continues on next page)

(continued from previous page)

```
Best pipeline AUC: 0.853164
```

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          0.853164              0.010011          0.863894
1          0.827102              0.027002          0.853667
2          0.826675              0.027444          0.853544
3          0.500000              0.000000          0.500000

      percent_better_than_baseline  high_variance_cv  \
0                35.316397              False
1                32.710221              False
2                32.667497              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', 0.8644456773933219),
                  ('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', 0.8097111537038906),
                  ('Fraud Cost', 0.0015780080694855455)])
```

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

(continues on next page)

(continued from previous page)

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']
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)

      automl.search()

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

(continues on next page)

(continued from previous page)

```
*****
* 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: catboost, linear_model, 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 + SMOTENC Oversampler + Standard_
↳Scaler:
```

```
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.014
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Elastic Net Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +
```

```
↳Standard Scaler may not perform as estimated on unseen data.
```

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

```
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.015
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Random Forest Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler
```

```
↳may not perform as estimated on unseen data.
```

```
Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +
↳Standard Scaler:
```

```
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.014
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC
```

```
↳Oversampler + Standard Scaler may not perform as estimated on unseen data.
```

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

```
  Starting cross validation
  Finished cross validation - mean Lead Scoring: 0.358
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  CatBoost Classifier w/ Imputer + SMOTENC Oversampler may not perform as
```

```
↳estimated on unseen data.
```

```
Search finished after 00:14
```

```
Best pipeline: CatBoost Classifier w/ Imputer + SMOTENC 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 + SMOTENC Overs...	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	0.015067	0.040868	-0.016129	
2	0.013992	0.041501	-0.016129	
3	0.013992	0.041501	-0.016129	
4	0.000000	0.000000	0.000000	

	percent_better_than_baseline	high_variance_cv	\
0	inf	True	
1	inf	True	
2	inf	True	
3	inf	True	
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 + SMOTENC Oversampler *
*****

Problem Type: binary
Model Family: CatBoost

Pipeline Steps
=====
1. Imputer

```

(continues on next page)

(continued from previous page)

```

    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. SMOTENC 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.

Objective to optimize binary classification pipeline thresholds for: <evalml.

↪objectives.lead_scoring.LeadScoring object at 0x7f86ba5f0760>

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)

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.
Allowed model families: catboost, linear_model, 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 + SMOTENC Oversampler + Standard_
↪Scaler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.653
Random Forest Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.652
Logistic Regression Classifier w/ Imputer + One Hot Encoder + SMOTENC Oversampler +_
↪Standard Scaler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.653
CatBoost Classifier w/ Imputer + SMOTENC Oversampler:
  Starting cross validation
  Finished cross validation - mean AUC: 0.882
```

(continues on next page)

(continued from previous page)

```

Search finished after 00:10
Best pipeline: CatBoost Classifier w/ Imputer + SMOTENC Oversampler
Best pipeline AUC: 0.881709

```

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

```

[12]:   id      pipeline_name  search_order  \
0    4  CatBoost Classifier w/ Imputer + SMOTENC Overs...      4
1    3  Logistic Regression Classifier w/ Imputer + On...      3
2    1  Elastic Net Classifier w/ Imputer + One Hot En...      1
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          0.653038              0.035883          0.691832
2          0.652855              0.035875          0.691582
3          0.651591              0.059622          0.717764
4          0.500000              0.000000          0.500000

      percent_better_than_baseline  high_variance_cv  \
0                38.170947              False
1                15.303770              False
2                15.285487              False
3                15.159098              False
4                 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      {'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,
                                              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)
      automl.search()

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

```

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

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

```

(continues on next page)

(continued from previous page)

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

Using SequentialEngine to train and score pipelines.
Searching up to 5 pipelines.
Allowed model families: decision_tree, linear_model, extra_trees, catboost, 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 Log Loss Binary: 9.164
```

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

```
Elastic Net Classifier w/ Imputer + One Hot Encoder + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.424
Decision Tree Classifier w/ Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.628
Random Forest Classifier w/ Imputer + One Hot Encoder:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.426
LightGBM Classifier w/ 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/ Imputer + One Hot Encoder + Standard Scaler
Best pipeline Log Loss Binary: 0.424184
```

```
[3]: OrderedDict([('Log Loss Binary', 0.4193730950794927)])
```

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', 53.43114055844771)])
```

```
[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)
```

```
376315.5229531472
```

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)
automl.search()

cbm_pipeline = automl.best_pipeline
```

Generating pipelines to search over...
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: decision_tree, linear_model, extra_trees, catboost, 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 Cost Benefit Matrix: -53.063

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

Elastic Net Classifier w/ Imputer + One Hot Encoder + Standard Scaler:
Starting cross validation
Finished cross validation - mean Cost Benefit Matrix: 59.392
Decision Tree Classifier w/ Imputer + One Hot Encoder:
Starting cross validation
Finished cross validation - mean Cost Benefit Matrix: 52.982
Random Forest Classifier w/ Imputer + One Hot Encoder:
Starting cross validation
Finished cross validation - mean Cost Benefit Matrix: 59.048
LightGBM Classifier w/ Imputer + One Hot Encoder:
Starting cross validation
Finished cross validation - mean Cost Benefit Matrix: 50.692
High coefficient of variation (cv >= 0.2) within cross validation scores.
LightGBM Classifier w/ Imputer + One Hot Encoder may not perform as estimated
 ↪ on unseen data.

Search finished after 00:14
Best pipeline: Elastic Net Classifier w/ Imputer + One Hot Encoder + Standard Scaler
Best pipeline Cost Benefit Matrix: 59.391831

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', 61.57122574538572)])

```
[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)
```

433646.1429247516

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

57330.619971604436

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
```

	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)

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.
Allowed model families: extra_trees, decision_tree, linear_model, random_forest,
    ↪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: 14.046

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

(continues on next page)

(continued from previous page)

```

Elastic Net Classifier w/ Text Featurization Component + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.350
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Elastic Net Classifier w/ Text Featurization Component + Standard Scaler may
    ↪not perform as estimated on unseen data.
Decision Tree Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 3.386
Random Forest Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.221
LightGBM Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.292
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    LightGBM Classifier w/ Text Featurization Component may not perform as
    ↪estimated on unseen data.
Logistic Regression Classifier w/ Text Featurization Component + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.350
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Logistic Regression Classifier w/ Text Featurization Component + Standard
    ↪Scaler may not perform as estimated on unseen data.
[14:30:34] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
    ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
    ↪behavior.
[14:30:35] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
    ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
    ↪behavior.
[14:30:36] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
    ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
    ↪behavior.
XGBoost Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.266
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    XGBoost Classifier w/ Text Featurization Component may not perform as
    ↪estimated on unseen data.
Extra Trees Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.292
CatBoost Classifier w/ Text Featurization Component:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.589

Search finished after 00:25
Best pipeline: Random Forest Classifier w/ Text Featurization Component
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	True	
2	97.922737	True	
3	97.918427	False	
4	97.505728	True	
5	97.504795	True	
6	95.806967	False	
7	75.896294	False	
8	0.000000	False	


```

parameters
0 {'Random Forest Classifier': {'n_estimators': ...
1 {'XGBoost Classifier': {'eta': 0.1, 'max_depth...
2 {'LightGBM Classifier': {'boosting_type': 'gbd...
3 {'Extra Trees Classifier': {'n_estimators': 10...
4 {'Logistic Regression Classifier': {'penalty':...
5 {'Elastic Net Classifier': {'penalty': 'elasti...
6 {'CatBoost Classifier': {'n_estimators': 10, '...
7 {'Decision Tree Classifier': {'criterion': 'gi...
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 *
*****

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Text Featurization Component
2. Random Forest Classifier
   * n_estimators : 100
   * max_depth : 6
   * n_jobs : -1

Training
=====
Training for binary problems.
Total training time (including CV): 3.0 seconds

Cross Validation
-----
```

	Log Loss Binary	MCC Binary	AUC	Precision	F1	Balanced Accuracy
↪ Binary Accuracy	Binary # Training	Binary # Validation				
0	0.222	0.817	0.975	0.862	0.893	0.
↪ 913	0.910	400	200			
1	0.180	0.875	0.985	0.937	0.925	0.
↪ 936	0.940	400	200			
2	0.262	0.783	0.963	0.918	0.865	0.
↪ 883	0.895	400	200			
mean	0.221	0.825	0.974	0.906	0.894	0.
↪ 910	0.915	-	-			
std	0.041	0.047	0.011	0.039	0.030	0.
↪ 026	0.023	-	-			
coef of var	0.185	0.057	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 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`.

```
[13]: 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)
```

```
[14]: automl_no_text = AutoMLSearch(X_train=X_train, y_train=y_train,
      problem_type='binary',
      max_batches=1,
      optimize_thresholds=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: extra_trees, decision_tree, linear_model, random_forest,
↪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: 14.046
```

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

(continues on next page)

(continued from previous page)

```

*****

Elastic Net Classifier w/ Text Featurization Component + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Elastic Net Classifier w/ Text Featurization Component + Standard Scaler may
  ↪not perform as estimated on unseen data.
Decision Tree Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 3.386
Random Forest Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.221
LightGBM Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  LightGBM Classifier w/ Text Featurization Component may not perform as
  ↪estimated on unseen data.
Logistic Regression Classifier w/ Text Featurization Component + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.350
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Text Featurization Component + Standard
  ↪Scaler may not perform as estimated on unseen data.
[14:31:00] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
  ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
  ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
  ↪behavior.
[14:31:01] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
  ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
  ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
  ↪behavior.
[14:31:02] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default
  ↪evaluation metric used with the objective 'binary:logistic' was changed from 'error
  ↪' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old
  ↪behavior.
XGBoost Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.266
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  XGBoost Classifier w/ Text Featurization Component may not perform as
  ↪estimated on unseen data.
Extra Trees Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.292
CatBoost Classifier w/ Text Featurization Component:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.589

Search finished after 00:23
Best pipeline: Random Forest Classifier w/ Text Featurization Component
Best pipeline Log Loss Binary: 0.221422

```

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

```
[15]: automl_no_text.rankings
```

```
[15]:
```

	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	True	
2	97.922737	True	
3	97.918427	False	
4	97.505728	True	
5	97.504795	True	
6	95.806967	False	
7	75.896294	False	
8	0.000000	False	


```

parameters
0 {'Random Forest Classifier': {'n_estimators': ...
1 {'XGBoost Classifier': {'eta': 0.1, 'max_depth...
2 {'LightGBM Classifier': {'boosting_type': 'gbd...
3 {'Extra Trees Classifier': {'n_estimators': 10...
4 {'Logistic Regression Classifier': {'penalty':...
5 {'Elastic Net Classifier': {'penalty': 'elasti...
6 {'CatBoost Classifier': {'n_estimators': 10, '...
7 {'Decision Tree Classifier': {'criterion': 'gi...
8 {'Baseline Classifier': {'strategy': 'mode'}}

```

```
[16]: 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.

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

```
[17]:
```

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

```
*****
```

(continues on next page)

(continued from previous page)

```

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

Problem Type: binary
Model Family: Random Forest

Pipeline Steps
=====
1. Text Featurization Component
2. Random Forest Classifier
   * n_estimators : 100
   * max_depth : 6
   * n_jobs : -1

Training
=====
Training for binary problems.
Total training time (including CV): 3.1 seconds

Cross Validation
-----

```

	Log Loss Binary	MCC Binary	AUC	Precision	F1	Balanced Accuracy
→ Binary	Accuracy Binary	# Training	# Validation			
0	0.222	0.817	0.975	0.862	0.893	0.
→ 913	0.910	400	200			
1	0.180	0.875	0.985	0.937	0.925	0.
→ 936	0.940	400	200			
2	0.262	0.783	0.963	0.918	0.865	0.
→ 883	0.895	400	200			
mean	0.221	0.825	0.974	0.906	0.894	0.
→ 910	0.915	-	-			
std	0.041	0.047	0.011	0.039	0.030	0.
→ 026	0.023	-	-			
coef of var	0.185	0.057	0.012	0.043	0.034	0.
→ 029	0.025	-	-			

```

[19]: # 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 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')
automl.search()

Using default limit of max_batches=1.

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, lightgbm, decision_tree, random_forest,
    ↪ xgboost, 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: 3.970

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

Elastic Net Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↪ Encoder + SMOTENC Oversampler + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.512
Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↪ Encoder + SMOTENC Oversampler: (continues on next page)
```

(continued from previous page)

```

    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 2.957
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One_
    ↳Hot Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.
Random Forest Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↳Encoder + SMOTENC Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.286
LightGBM Classifier w/ Imputer + DateTime Featurization Component + One Hot Encoder +_
    ↳SMOTENC Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.309
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    LightGBM Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↳Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.
Logistic Regression Classifier w/ Imputer + DateTime Featurization Component + One_
    ↳Hot Encoder + SMOTENC Oversampler + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.552
[14:32:02] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default_
    ↳evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↳' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old_
    ↳behavior.
[14:32:03] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default_
    ↳evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↳' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old_
    ↳behavior.
[14:32:04] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default_
    ↳evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↳' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old_
    ↳behavior.
XGBoost Classifier w/ Imputer + DateTime Featurization Component + One Hot Encoder +_
    ↳SMOTENC Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.279
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    XGBoost Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↳Encoder + SMOTENC Oversampler may not perform as estimated on unseen data.
Extra Trees Classifier w/ Imputer + DateTime Featurization Component + One Hot_
    ↳Encoder + SMOTENC Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.338
CatBoost Classifier w/ Imputer + DateTime Featurization Component + SMOTENC_
    ↳Oversampler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.602

Search finished after 00:19
[14:32:08] WARNING: ../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default_
    ↳evaluation metric used with the objective 'binary:logistic' was changed from 'error
    ↳' to 'logloss'. Explicitly set eval_metric if you'd like to restore the old_
    ↳behavior.
Best pipeline: XGBoost Classifier w/ Imputer + DateTime Featurization Component + One_
    ↳Hot Encoder + SMOTENC Oversampler
Best pipeline Log Loss Binary: 0.278707

```


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',
                                           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.
```

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/ 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
5   5  Logistic Regression Classifier w/ Imputer + Da...    5
6   8  CatBoost Classifier w/ Imputer + DateTime Feat...    8
```

(continues on next page)

(continued from previous page)

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	0.278707	0.190725	0.202638
1	0.285613	0.041116	0.263695
2	0.308636	0.203878	0.234947
3	0.338286	0.009381	0.329015
4	0.511808	0.074992	0.590517
5	0.551960	0.082695	0.628646
6	0.601819	0.007246	0.593693
7	2.956956	3.084439	0.651557
8	3.970423	0.266060	4.124033
	percent_better_than_baseline	high_variance_cv \	
0	92.980409	True	
1	92.806475	False	
2	92.226623	True	
3	91.479857	False	
4	87.109474	False	
5	86.098206	False	
6	84.842444	False	
7	25.525413	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' }}		

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/ Imputer + DateTime Featurization Component + One Hot_
↳Encoder + SMOTENC Oversampler + Standard Scaler *
*****

Problem Type: binary
Model Family: Linear

Pipeline Steps
=====
1. Imputer
```

(continues on next page)

(continued from previous page)

```

    * categorical_impute_strategy : most_frequent
    * numeric_impute_strategy : mean
    * categorical_fill_value : None
    * numeric_fill_value : None
2. DateTime Featurization Component
    * features_to_extract : ['year', 'month', 'day_of_week', 'hour']
    * encode_as_categories : False
    * date_index : None
3. One Hot Encoder
    * top_n : 10
    * features_to_encode : None
    * categories : None
    * drop : if_binary
    * handle_unknown : ignore
    * handle_missing : error
4. SMOTENC Oversampler
    * sampling_ratio : 0.25
    * k_neighbors_default : 5
    * n_jobs : -1
    * sampling_ratio_dict : None
    * categorical_features : [3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
↪22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
↪43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59]
    * k_neighbors : 5
5. Standard Scaler
6. 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): 2.2 seconds

Cross Validation

```

-----
          Log Loss Binary  MCC Binary  AUC  Precision  F1  Balanced Accuracy_
↪Binary  Accuracy Binary # Training # Validation
0          0.591          133  0.175 0.597          0.286 0.267          0.
↪583          0.836          133          67
1          0.441          133  0.296 0.710          0.500 0.333          0.
↪608          0.881          133          67
2          0.504          134  0.035 0.630          0.120 0.188          0.
↪528          0.606          134          66
mean          0.512          -  0.169 0.646          0.302 0.263          0.
↪573          0.774          -          -          0.191 0.073          0.
std          0.075          -  0.130 0.058          0.191 0.073          0.
↪041          0.147          -          -          0.631 0.278          0.
coef of var          0.147          -  0.772 0.090          0.631 0.278          0.
↪072          0.190          -          -

```

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/ Imputer + DateTime Featurization Component + One Hot
↳Encoder + SMOTENC Oversampler + Standard Scaler
{'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'}, 'SMOTENC Oversampler': {
↳'sampling_ratio': 0.25, 'k_neighbors_default': 5, 'n_jobs': -1, 'sampling_ratio_dict
↳': None, 'categorical_features': [3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
↳ 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
↳ 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59], '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/ Imputer + DateTime Featurization Component + One Hot Encoder +
↳SMOTENC Oversampler
{'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'}, 'SMOTENC Oversampler': {
↳'sampling_ratio': 0.25, 'k_neighbors_default': 5, 'n_jobs': -1, 'sampling_ratio_dict
↳': None, 'categorical_features': [3, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,
↳ 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
↳ 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59], 'k_neighbors':
↳ 5}, 'XGBoost Classifier': {'eta': 0.1, 'max_depth': 6, 'min_child_weight': 1, 'n_
↳estimators': 100, 'n_jobs': -1}}
```

```
[15]: 0      False
      1      False
      2      False
      3       True
      4      False
      ...
     195     False
     196     False
```

(continues on next page)

(continued from previous page)

```
[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/ Imputer + DateTime Featurization Component + One Hot_
↳ Encoder + SMOTENC Oversampler + Standard Scaler': OrderedDict([('Accuracy Binary',
0.66),
('F1', 0.19047619047619044),
('AUC', 0.5265151515151515)]),
'Decision Tree Classifier w/ Imputer + DateTime Featurization Component + One Hot_
↳ Encoder + SMOTENC Oversampler': OrderedDict([('Accuracy Binary',
0.92),
('F1', 0.6),
('AUC', 0.7234848484848486)])}
```

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 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)
```

	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]

Using default limit of max_batches=1.

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

[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'], 'SMOTENC Oversampler': ['SMOTENC Oversampler', 'One Hot Encoder.
    ↳ x', 'y'], 'Standard Scaler': ['Standard Scaler', 'SMOTENC Oversampler.x', 'SMOTENC_
    ↳ Oversampler.y'], 'Logistic Regression Classifier': ['Logistic Regression Classifier
    ↳ ', 'Standard Scaler.x', 'SMOTENC 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'}, 'SMOTENC 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 `SMOTENC 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.

```
[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]

Using default limit of max_batches=1.

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

[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)
```

(continued from previous page)

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 majority
sampler_ratio_dict = {0: 1, 1: 0.5}
pipeline_parameters = {"SMOTENC 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 = {"SMOTENC 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)
```

Using default limit of `max_batches=1`.

Generating pipelines to search over...

8 pipelines ready for search.

```
[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'], 'SMOTENC Oversampler': ['SMOTENC Oversampler', 'One Hot Encoder.
# x', 'y'], 'Standard Scaler': ['Standard Scaler', 'SMOTENC Oversampler.x', 'SMOTENC
# Oversampler.y'], 'Logistic Regression Classifier': ['Logistic Regression Classifier
# ', 'Standard Scaler.x', 'SMOTENC 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'}, 'SMOTENC 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 algorithm 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)

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
```

(continues on next page)

(continued from previous page)

```

*****
* 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
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + Standard Scaler may not perform
↳as estimated on unseen data.

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

Logistic Regression Classifier w/ Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.097
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + Standard Scaler may not perform
↳as estimated on unseen data.
Logistic Regression Classifier w/ Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.085
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + Standard Scaler may not perform
↳as estimated on unseen data.
Logistic Regression Classifier w/ Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.097
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + Standard Scaler may not perform
↳as estimated on unseen data.
Logistic Regression Classifier w/ Imputer + Standard Scaler:
  Starting cross validation
  Finished cross validation - mean Log Loss Binary: 0.091
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Logistic Regression Classifier w/ Imputer + Standard Scaler may not perform
↳as estimated on unseen data.
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
  High coefficient of variation (cv >= 0.2) within cross validation scores.
  Elastic Net Classifier w/ Imputer + Standard Scaler may not perform as
↳estimated on unseen data.
Elastic Net Classifier w/ Imputer + Standard Scaler:

```

(continues on next page)

(continued from previous page)

```

    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.075
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Elastic Net Classifier w/ Imputer + Standard Scaler may not perform as
    ↪estimated on unseen data.
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
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Elastic Net Classifier w/ Imputer + Standard Scaler may not perform as
    ↪estimated on unseen data.
Elastic Net Classifier w/ Imputer + Standard Scaler:
    Starting cross validation
    Finished cross validation - mean Log Loss Binary: 0.075
    High coefficient of variation (cv >= 0.2) within cross validation scores.
    Elastic Net Classifier w/ Imputer + Standard Scaler may not perform as
    ↪estimated on unseen data.

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

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

Search finished after 00:23
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

```

(continues on next page)

(continued from previous page)

```
* C : 8.123565600467177
* ll_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': 0.8667762279510498,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
        4.124033002377396),
        ('MCC Binary', 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',
        4.124033002377395),
        ('MCC Binary', 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),
        ('AUC', 0.5),
        ('Precision', 0.0),
        ('F1', 0.0),
        ('Balanced Accuracy Binary', 0.5),
        ('Accuracy Binary', 0.8939393939393939),
```

(continues on next page)

(continued from previous page)

```

        ('# 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,
    '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/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler + Standard Scaler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'Elastic Net Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler + Standard Scaler',
    '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'},
    'SMOTENC Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'categorical_features': [3,
    10,
    11,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    21,
    22,
    23,
    24,
    25,

```

(continues on next page)

(continued from previous page)

```

26,
27,
28,
29,
30,
31,
32,
33,
34,
35,
36,
37,
38,
39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
57,
58,
59],
'k_neighbors': 5},
'Elastic Net Classifier': {'penalty': 'elasticnet',
'C': 1.0,
'11_ratio': 0.15,
'n_jobs': -1,
'multi_class': 'auto',
'solver': 'saga'}},
'mean_cv_score': 0.5118084226551022,
'standard_deviation_cv_score': 0.07499222153032764,
'high_variance_cv': False,
'training_time': 2.2299022674560547,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.590516914375052),
('MCC Binary', 0.17518582316850065),
('AUC', 0.597457627118644),
('Precision', 0.2857142857142857),
('F1', 0.26666666666666666),
('Balanced Accuracy Binary', 0.5826271186440678),
('Accuracy Binary', 0.835820895522388),
('# Training', 133),
('# Validation', 67)])},
'mean_cv_score': 0.590516914375052,
'binary_classification_threshold': 0.48089822234602875},

```

(continues on next page)

(continued from previous page)

```

{'all_objective_scores': OrderedDict([('Log Loss Binary',
                                     0.44118816828058105),
                                     ('MCC Binary', 0.2957528252716689),
                                     ('AUC', 0.7097457627118644),
                                     ('Precision', 0.5),
                                     ('F1', 0.3333333333333333),
                                     ('Balanced Accuracy Binary', 0.6080508474576272),
                                     ('Accuracy Binary', 0.8805970149253731),
                                     ('# Training', 133),
                                     ('# Validation', 67)]),
 'mean_cv_score': 0.44118816828058105,
 'binary_classification_threshold': 0.7518116371406225},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
                                     0.5037201853096736),
                                     ('MCC Binary', 0.035350118786872956),
                                     ('AUC', 0.6295399515738498),
                                     ('Precision', 0.12),
                                     ('F1', 0.1875),
                                     ('Balanced Accuracy Binary', 0.5278450363196125),
                                     ('Accuracy Binary', 0.6060606060606061),
                                     ('# Training', 134),
                                     ('# Validation', 66)]),
 'mean_cv_score': 0.5037201853096736,
 'binary_classification_threshold': 0.10568708806540947}},
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 87.
↪10947426720426,
'MCC Binary': inf,
'AUC': 14.558111380145277,
'Precision': 30.19047619047619,
'F1': 26.25,
'Balanced Accuracy Binary': 7.284100080710254,
'Accuracy Binary': -11.08849690939243},
'percent_better_than_baseline': 87.10947426720426,
'validation_score': 0.590516914375052},
2: {'id': 2,
    'pipeline_name': 'Decision Tree Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'Decision Tree Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
    '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'}},

```

(continues on next page)

(continued from previous page)

```
'SMOTENC Oversampler': {'sampling_ratio': 0.25,
  'k_neighbors_default': 5,
  'n_jobs': -1,
  'sampling_ratio_dict': None,
  'categorical_features': [3,
    10,
    11,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    21,
    22,
    23,
    24,
    25,
    26,
    27,
    28,
    29,
    30,
    31,
    32,
    33,
    34,
    35,
    36,
    37,
    38,
    39,
    40,
    41,
    42,
    43,
    44,
    45,
    46,
    47,
    48,
    49,
    50,
    51,
    52,
    53,
    54,
    55,
    56,
    57,
    58,
    59],
  'k_neighbors': 5},
'Decision Tree Classifier': {'criterion': 'gini',
```

(continues on next page)

(continued from previous page)

```

    'max_features': 'auto',
    'max_depth': 6,
    'min_samples_split': 2,
    'min_weight_fraction_leaf': 0.0}},
    'mean_cv_score': 2.956956288246977,
    'standard_deviation_cv_score': 3.084438905133218,
    'high_variance_cv': True,
    'training_time': 1.8792915344238281,
    'cv_data': [{ 'all_objective_scores': OrderedDict([('Log Loss Binary',
                                                         0.651556943513563),
                                                         ('MCC Binary', 0.7712055916006633),
                                                         ('AUC', 0.8103813559322034),
                                                         ('Precision', 1.0),
                                                         ('F1', 0.7692307692307693),
                                                         ('Balanced Accuracy Binary', 0.8125),
                                                         ('Accuracy Binary', 0.9552238805970149),
                                                         ('# Training', 133),
                                                         ('# Validation', 67)]),
                  'mean_cv_score': 0.651556943513563,
                  'binary_classification_threshold': 0.9999999885588493},
                 { 'all_objective_scores': OrderedDict([('Log Loss Binary',
                                                         1.7585681792927779),
                                                         ('MCC Binary', 0.4208952550769721),
                                                         ('AUC', 0.6896186440677965),
                                                         ('Precision', 0.6),
                                                         ('F1', 0.4615384615384615),
                                                         ('Balanced Accuracy Binary', 0.6705508474576272),
                                                         ('Accuracy Binary', 0.8955223880597015),
                                                         ('# Training', 133),
                                                         ('# Validation', 67)]),
                  'mean_cv_score': 1.7585681792927779,
                  'binary_classification_threshold': 0.9999999885588493},
                 { 'all_objective_scores': OrderedDict([('Log Loss Binary',
                                                         6.460743741934591),
                                                         ('MCC Binary', -0.20116515012026231),
                                                         ('AUC', 0.3523002421307506),
                                                         ('Precision', 0.06521739130434782),
                                                         ('F1', 0.11320754716981131),
                                                         ('Balanced Accuracy Binary', 0.3498789346246973),
                                                         ('Accuracy Binary', 0.2878787878787879),
                                                         ('# Training', 134),
                                                         ('# Validation', 66)]),
                  'mean_cv_score': 6.460743741934591,
                  'binary_classification_threshold': 0.1568627424310231}}],
    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 25.
↪52541256225873,
    'MCC Binary': inf,
    'AUC': 11.743341404358354,
    'Precision': 55.5072463768116,
    'F1': 44.79922593130141,
    'Balanced Accuracy Binary': 11.097659402744153,
    'Accuracy Binary': -17.21694557515452},
    'percent_better_than_baseline': 25.52541256225873,
    'validation_score': 0.651556943513563},
    3: {'id': 3,
        'pipeline_name': 'Random Forest Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',

```

(continues on next page)

(continued from previous page)

```

'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
'pipeline_summary': 'Random Forest Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
'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'},
'SMOTENC Oversampler': {'sampling_ratio': 0.25,
'k_neighbors_default': 5,
'n_jobs': -1,
'sampling_ratio_dict': None,
'categorical_features': [3,
10,
11,
12,
13,
14,
15,
16,
17,
18,
19,
20,
21,
22,
23,
24,
25,
26,
27,
28,
29,
30,
31,
32,
33,
34,
35,
36,
37,
38,
39,
40,
41,

```

(continues on next page)

(continued from previous page)

```

42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
57,
58,
59],
'k_neighbors': 5},
'Random Forest Classifier': {'n_estimators': 100,
'max_depth': 6,
'n_jobs': -1}},
'mean_cv_score': 0.2856133806139974,
'standard_deviation_cv_score': 0.041115539258178,
'high_variance_cv': False,
'training_time': 2.3439459800720215,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.26369520661129536),
('MCC Binary', 0.47636443708895493),
('AUC', 0.8241525423728814),
('Precision', 1.0),
('F1', 0.4),
('Balanced Accuracy Binary', 0.625),
('Accuracy Binary', 0.9104477611940298),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.26369520661129536,
'binary_classification_threshold': 0.4988241268809967},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.33304413887923806),
('MCC Binary', 0.4900218379501181),
('AUC', 0.711864406779661),
('Precision', 0.75),
('F1', 0.5),
('Balanced Accuracy Binary', 0.6790254237288136),
('Accuracy Binary', 0.9104477611940298),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.33304413887923806,
'binary_classification_threshold': 0.5065550313835498},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.2601007963514588),
('MCC Binary', 0.6335302236023843),
('AUC', 0.8789346246973365),
('Precision', 1.0),
('F1', 0.6),
('Balanced Accuracy Binary', 0.7142857142857143),

```

(continues on next page)

(continued from previous page)

```

        ('Accuracy Binary', 0.9393939393939394),
        ('# Training', 134),
        ('# Validation', 66)]),
    'mean_cv_score': 0.2601007963514588,
    'binary_classification_threshold': 0.4459844896470452]],
    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 92.
↪80647509992905,
    'MCC Binary': inf,
    'AUC': 30.498385794995965,
    'Precision': 91.66666666666666,
    'F1': 50.0,
    'Balanced Accuracy Binary': 17.277037933817596,
    'Accuracy Binary': 3.505201266395308},
    'percent_better_than_baseline': 92.80647509992905,
    'validation_score': 0.26369520661129536},
4: {'id': 4,
    'pipeline_name': 'LightGBM Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'LightGBM Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
    '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'},
    'SMOTENC Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'categorical_features': [3,
    10,
    11,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    21,
    22,
    23,

```

(continues on next page)

(continued from previous page)

```

24,
25,
26,
27,
28,
29,
30,
31,
32,
33,
34,
35,
36,
37,
38,
39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
57,
58,
59],
'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': 0.30863594948686357,
'standard_deviation_cv_score': 0.20387814963979614,
'high_variance_cv': True,
'training_time': 1.9574458599090576,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.23494740261438743),
('MCC Binary', 0.4900218379501181),
('AUC', 0.8940677966101696),
('Precision', 0.75),
('F1', 0.5),
('Balanced Accuracy Binary', 0.6790254237288136),

```

(continues on next page)

(continued from previous page)

```

        ('Accuracy Binary', 0.9104477611940298),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.23494740261438743,
    'binary_classification_threshold': 0.5037478165498673},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
        0.5391133772263208),
        ('MCC Binary', 0.4900218379501181),
        ('AUC', 0.635593220338983),
        ('Precision', 0.75),
        ('F1', 0.5),
        ('Balanced Accuracy Binary', 0.6790254237288136),
        ('Accuracy Binary', 0.9104477611940298),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.5391133772263208,
    'binary_classification_threshold': 0.7736716431132877},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
        0.15184706861988254),
        ('MCC Binary', 0.3600976668493281),
        ('AUC', 0.8934624697336562),
        ('Precision', 1.0),
        ('F1', 0.25),
        ('Balanced Accuracy Binary', 0.5714285714285714),
        ('Accuracy Binary', 0.9090909090909091),
        ('# Training', 134),
        ('# Validation', 66)]),
    'mean_cv_score': 0.15184706861988254,
    'binary_classification_threshold': 0.9385567379765596}],
    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 92.
↪22662333635083,
    'MCC Binary': inf,
    'AUC': 30.77078288942697,
    'Precision': 83.33333333333334,
    'F1': 41.66666666666667,
    'Balanced Accuracy Binary': 14.315980629539949,
    'Accuracy Binary': 2.4951002562942914},
    'percent_better_than_baseline': 92.22662333635083,
    'validation_score': 0.23494740261438743},
    5: {'id': 5,
    'pipeline_name': 'Logistic Regression Classifier w/ Imputer + DateTime_
↪Featurization Component + One Hot Encoder + SMOTENC Oversampler + Standard Scaler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'Logistic Regression Classifier w/ Imputer + DateTime_
↪Featurization Component + One Hot Encoder + SMOTENC Oversampler + Standard Scaler',
    '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,

```

(continues on next page)

(continued from previous page)

```
'features_to_encode': None,
'categories': None,
'drop': 'if_binary',
'handle_unknown': 'ignore',
'handle_missing': 'error'},
'SMOTENC Oversampler': {'sampling_ratio': 0.25,
'k_neighbors_default': 5,
'n_jobs': -1,
'sampling_ratio_dict': None,
'categorical_features': [3,
10,
11,
12,
13,
14,
15,
16,
17,
18,
19,
20,
21,
22,
23,
24,
25,
26,
27,
28,
29,
30,
31,
32,
33,
34,
35,
36,
37,
38,
39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
```

(continues on next page)

(continued from previous page)

```

57,
58,
59],
'k_neighbors': 5},
'Logistic Regression Classifier': {'penalty': 'l2',
'C': 1.0,
'n_jobs': -1,
'multi_class': 'auto',
'solver': 'lbfgs'}},
'mean_cv_score': 0.5519600330274397,
'standard_deviation_cv_score': 0.08269486722266915,
'high_variance_cv': False,
'training_time': 3.6108527183532715,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.6286458411705291),
('MCC Binary', 0.17518582316850065),
('AUC', 0.5826271186440678),
('Precision', 0.2857142857142857),
('F1', 0.26666666666666666),
('Balanced Accuracy Binary', 0.5826271186440678),
('Accuracy Binary', 0.835820895522388),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.6286458411705291,
'binary_classification_threshold': 0.4313366338786227},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.46434334092355634),
('MCC Binary', 0.14283901342792224),
('AUC', 0.7033898305084746),
('Precision', 0.3333333333333333),
('F1', 0.18181818181818182),
('Balanced Accuracy Binary', 0.5455508474576272),
('Accuracy Binary', 0.8656716417910447),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.46434334092355634,
'binary_classification_threshold': 0.7298626702409028},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.5628909169882338),
('MCC Binary', 0.13468820490061437),
('AUC', 0.6077481840193704),
('Precision', 0.17647058823529413),
('F1', 0.25),
('Balanced Accuracy Binary', 0.5956416464891041),
('Accuracy Binary', 0.7272727272727273),
('# Training', 134),
('# Validation', 66)]),
'mean_cv_score': 0.5628909169882338,
'binary_classification_threshold': 0.16920770831511384}],
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 86.
↪09820648846629,
'MCC Binary': inf,
'AUC': 13.125504439063763,
'Precision': 26.517273576097107,
'F1': 23.282828282828284,
'Balanced Accuracy Binary': 7.4606537530266355,
'Accuracy Binary': -7.5456053067993185},

```

(continues on next page)

(continued from previous page)

```

    'percent_better_than_baseline': 86.09820648846629,
    'validation_score': 0.6286458411705291}},
6: {'id': 6,
    'pipeline_name': 'XGBoost Classifier w/ Imputer + DateTime Featurization Component_
↪+ One Hot Encoder + SMOTENC Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'XGBoost Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
    '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'},
    'SMOTENC Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'categorical_features': [3,
    10,
    11,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    21,
    22,
    23,
    24,
    25,
    26,
    27,
    28,
    29,
    30,
    31,
    32,
    33,
    34,
    35,
    36,

```

(continues on next page)

(continued from previous page)

```

37,
38,
39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
57,
58,
59],
'k_neighbors': 5},
'XGBoost Classifier': {'eta': 0.1,
'max_depth': 6,
'min_child_weight': 1,
'n_estimators': 100,
'n_jobs': -1}},
'mean_cv_score': 0.2787074593034545,
'standard_deviation_cv_score': 0.19072502137726743,
'high_variance_cv': True,
'training_time': 2.0405526161193848,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.20263786688400182),
('MCC Binary', 0.6842908506285673),
('AUC', 0.8919491525423728),
('Precision', 1.0),
('F1', 0.6666666666666666),
('Balanced Accuracy Binary', 0.75),
('Accuracy Binary', 0.9402985074626866),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.20263786688400182,
'binary_classification_threshold': 0.8503084678293262},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.4957285583724021),
('MCC Binary', 0.20588632450454833),
('AUC', 0.6483050847457628),
('Precision', 0.5),
('F1', 0.2),
('Balanced Accuracy Binary', 0.5540254237288136),
('Accuracy Binary', 0.8805970149253731),
('# Training', 133),
('# Validation', 67)]),
'mean_cv_score': 0.4957285583724021,
'binary_classification_threshold': 0.9268544884142358},

```

(continues on next page)

(continued from previous page)

```

{'all_objective_scores': OrderedDict([('Log Loss Binary',
                                     0.13775595265395954),
                                     ('MCC Binary', 0.3600976668493281),
                                     ('AUC', 0.8886198547215496),
                                     ('Precision', 1.0),
                                     ('F1', 0.25),
                                     ('Balanced Accuracy Binary', 0.5714285714285714),
                                     ('Accuracy Binary', 0.9090909090909091),
                                     ('# Training', 134),
                                     ('# Validation', 66)]),
 'mean_cv_score': 0.13775595265395954,
 'binary_classification_threshold': 0.9457477876739043}],
 'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 92.
↪98040923704308,
 'MCC Binary': inf,
 'AUC': 30.96246973365617,
 'Precision': 83.33333333333334,
 'F1': 37.22222222222222,
 'Balanced Accuracy Binary': 12.51513317191283,
 'Accuracy Binary': 2.4951002562942914},
 'percent_better_than_baseline': 92.98040923704308,
 'validation_score': 0.20263786688400182},
 7: {'id': 7,
     'pipeline_name': 'Extra Trees Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
     'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
     'pipeline_summary': 'Extra Trees Classifier w/ Imputer + DateTime Featurization_
↪Component + One Hot Encoder + SMOTENC Oversampler',
     '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'},
                     'SMOTENC Oversampler': {'sampling_ratio': 0.25,
                                         'k_neighbors_default': 5,
                                         'n_jobs': -1,
                                         'sampling_ratio_dict': None,
                                         'categorical_features': [3,
                                         10,
                                         11,
                                         12,
                                         13,
                                         14,
                                         15,
                                         16,

```

(continues on next page)

(continued from previous page)

```

17,
18,
19,
20,
21,
22,
23,
24,
25,
26,
27,
28,
29,
30,
31,
32,
33,
34,
35,
36,
37,
38,
39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
52,
53,
54,
55,
56,
57,
58,
59],
'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': 0.33828572013370833,
'standard_deviation_cv_score': 0.009380537483031018,
'high_variance_cv': False,
'training_time': 2.227367877960205,
'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
0.32901507195041363),
('MCC Binary', 0.3624510999859961),

```

(continues on next page)

(continued from previous page)

```

        ('AUC', 0.7838983050847458),
        ('Precision', 0.4),
        ('F1', 0.4444444444444445),
        ('Balanced Accuracy Binary', 0.6991525423728814),
        ('Accuracy Binary', 0.8507462686567164),
        ('# Training', 133),
        ('# Validation', 67)]),
    'mean_cv_score': 0.32901507195041363,
    'binary_classification_threshold': 0.24464061936922427},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.3380696737832299),
    ('MCC Binary', -0.045325960909933655),
    ('AUC', 0.7139830508474576),
    ('Precision', 0.0),
    ('F1', 0.0),
    ('Balanced Accuracy Binary', 0.4915254237288136),
    ('Accuracy Binary', 0.8656716417910447),
    ('# Training', 133),
    ('# Validation', 67)]),
    'mean_cv_score': 0.3380696737832299,
    'binary_classification_threshold': 0.3074199530907704},
{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.3477724146674814),
    ('MCC Binary', 0.27835560025568845),
    ('AUC', 0.7191283292978208),
    ('Precision', 0.22727272727272727),
    ('F1', 0.3448275862068965),
    ('Balanced Accuracy Binary', 0.7130750605326877),
    ('Accuracy Binary', 0.7121212121212122),
    ('# Training', 134),
    ('# Validation', 66)]),
    'mean_cv_score': 0.3477724146674814,
    'binary_classification_threshold': 0.2199210188616361}},
'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 91.
↪47985733060223,
    'MCC Binary': inf,
    'AUC': 23.900322841000797,
    'Precision': 20.90909090909091,
    'F1': 26.309067688378036,
    'Balanced Accuracy Binary': 13.458434221146087,
    'Accuracy Binary': -7.553143374038884},
'percent_better_than_baseline': 91.47985733060223,
'validation_score': 0.32901507195041363},
8: {'id': 8,
    'pipeline_name': 'CatBoost Classifier w/ Imputer + DateTime Featurization_
↪Component + SMOTENC Oversampler',
    'pipeline_class': evalml.pipelines.binary_classification_pipeline.
↪BinaryClassificationPipeline,
    'pipeline_summary': 'CatBoost Classifier w/ Imputer + DateTime Featurization_
↪Component + SMOTENC Oversampler',
    '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',

```

(continues on next page)

(continued from previous page)

```

    'hour'],
    'encode_as_categories': False,
    'date_index': None},
    'SMOTENC Oversampler': {'sampling_ratio': 0.25,
    'k_neighbors_default': 5,
    'n_jobs': -1,
    'sampling_ratio_dict': None,
    'categorical_features': [3, 4, 5, 6, 9, 10],
    'k_neighbors': 5},
    'CatBoost Classifier': {'n_estimators': 10,
    'eta': 0.03,
    'max_depth': 6,
    'bootstrap_type': None,
    'silent': True,
    'allow_writing_files': False,
    'n_jobs': -1}},
    'mean_cv_score': 0.6018191346985708,
    'standard_deviation_cv_score': 0.007246095254215931,
    'high_variance_cv': False,
    'training_time': 0.9588806629180908,
    'cv_data': [{'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.5936927052183711),
    ('MCC Binary', 0.7712055916006633),
    ('AUC', 0.8898305084745762),
    ('Precision', 1.0),
    ('F1', 0.7692307692307693),
    ('Balanced Accuracy Binary', 0.8125),
    ('Accuracy Binary', 0.9552238805970149),
    ('# Training', 133),
    ('# Validation', 67))]),
    'mean_cv_score': 0.5936927052183711,
    'binary_classification_threshold': 0.498623875482405},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.6076076767585368),
    ('MCC Binary', 0.47636443708895493),
    ('AUC', 0.5603813559322034),
    ('Precision', 1.0),
    ('F1', 0.4),
    ('Balanced Accuracy Binary', 0.625),
    ('Accuracy Binary', 0.9104477611940298),
    ('# Training', 133),
    ('# Validation', 67))]),
    'mean_cv_score': 0.6076076767585368,
    'binary_classification_threshold': 0.5007469614241519},
    {'all_objective_scores': OrderedDict([('Log Loss Binary',
    0.6041570221188047),
    ('MCC Binary', 0.7469064529073725),
    ('AUC', 0.9346246973365617),
    ('Precision', 0.8333333333333334),
    ('F1', 0.7692307692307692),
    ('Balanced Accuracy Binary', 0.8486682808716708),
    ('Accuracy Binary', 0.9545454545454546),
    ('# Training', 134),
    ('# Validation', 66))]),
    'mean_cv_score': 0.6041570221188047,
    'binary_classification_threshold': 0.47437131050216075}],
    'percent_better_than_baseline_all_objectives': {'Log Loss Binary': 84.
    ↪84244358059615,
```

(continues on next page)

(continued from previous page)

```
'MCC Binary': inf,
'AUC': 29.49455205811139,
'Precision': 94.44444444444446,
'F1': 64.61538461538461,
'Balanced Accuracy Binary': 26.205609362389026,
'Accuracy Binary': 5.502789084878645},
'percent_better_than_baseline': 84.84244358059615,
'validation_score': 0.5936927052183711}},
'search_order': [0, 1, 2, 3, 4, 5, 6, 7, 8]}
```

4.2 Objectives

4.2.1 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.

4.2.2 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
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

X, y = load_fraud(n_rows=100)
X.ww.init(logical_types={"provider": "Categorical", "region": "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)])
```

4.2.3 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.

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
```

(continues on next page)

(continued from previous page)

```

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

    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

```

(continues on next page)

(continued from previous page)

```

    # 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()

    return loss_per_total_processed

```

4.3 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**.

4.3.1 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()
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: 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 Transformer
Transformer: Drop Columns Transformer
Transformer: Undersampler
Transformer: SMOTEN Oversampler
Transformer: SMOTENC Oversampler
Transformer: SMOTE Oversampler
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: Polynomial Detrender

4.3.2 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: Stacked Ensemble Regressor
Estimator: 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
```


4.3.3 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', '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.

4.3.4 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

4.4 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.

4.4.1 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'],
↳ 'Random Forest Classifier': ['Random Forest Classifier', 'Imputer.x']}, 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'],
    'Encoder': ['One Hot Encoder', 'Imputer'],
    'Random Forest Clf': ['Random Forest Classifier', 'Encoder'],
    'Elastic Net Clf': ['Elastic Net Classifier', 'Encoder'],
    'Final Estimator': ['Logistic Regression Classifier', 'Random Forest Clf',
↳ 'Elastic Net Clf']
}

MulticlassClassificationPipeline(component_graph=component_graph_as_dict)
```

```
[2]: pipeline = MulticlassClassificationPipeline(component_graph={'Imputer': ['Imputer'],
↳ 'Encoder': ['One Hot Encoder', 'Imputer'], 'Random Forest Clf': ['Random Forest_
↳ Classifier', 'Encoder'], 'Elastic Net Clf': ['Elastic Net Classifier', 'Encoder'],
↳ 'Final Estimator': ['Logistic Regression Classifier', 'Random Forest Clf', 'Elastic_
↳ Net Clf']}, 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], 'Random Forest Classifier': ['Random Forest Classifier', 'New_
↳ Transformer.x']}, parameters={'New Transformer':{'parameter_1': 1}, 'Random Forest_
↳ Classifier':{'n_estimators': 100, 'max_depth': 6, 'n_jobs': -1}}, random_seed=0)
```

4.4.2 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
```

(continues on next page)

(continued from previous page)

```

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)])

```

4.4.3 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)

      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

```

4.4.4 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",

```

(continues on next page)

(continued from previous page)

```

        "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'],
↪'One Hot Encoder': ['One Hot Encoder', 'Imputer.x'], 'Standard Scaler': ['Standard_
↪Scaler', 'One Hot Encoder.x'], 'Logistic Regression Classifier': ['Logistic_
↪Regression Classifier', 'Standard Scaler.x']}, 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)

```

4.4.5 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'],
    'Encoder': ['One Hot Encoder', 'Imputer'],
    'Random Forest Clf': ['Random Forest Classifier', 'Encoder'],
    'Elastic Net Clf': ['Elastic Net Classifier', 'Encoder'],
    'Final Estimator': ['Logistic Regression Classifier', 'Random Forest Clf',
↪'Elastic Net Clf']
}

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 *

```

(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 : 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 *
*****

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

```

(continues on next page)

(continued from previous page)

```

* 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

```

4.4.6 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')

```

4.4.7 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')

```

4.4.8 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']}
```

4.4.9 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'],
↳ 'Random Forest Classifier': ['Random Forest Classifier', 'Imputer.x']}, 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)
```

4.4.10 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.
        """
        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:
```

(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): Targets
        Returns:
            pd.DataFrame: Transformed X
        """

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

pipeline_instance = MulticlassClassificationPipeline(['Imputer', MyDropNullColumns,
    ↪ 'DateTime Featurization_
    ↪ Component',
    ↪ 'Text Featurization Component',
    ↪ 'One Hot Encoder', 'Random_
    ↪ Forest Classifier'],
    ↪ custom_name="Pipeline with_
    ↪ 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'],
    ↪ 'My Drop Null Columns Transformer': [MyDropNullColumns, 'Imputer.x'], 'DateTime_
    ↪ Featurization Component': ['DateTime Featurization Component', 'My Drop Null_
    ↪ Columns Transformer.x'], 'Text Featurization Component': ['Text Featurization_
    ↪ Component', 'DateTime Featurization Component.x'], 'One Hot Encoder': ['One Hot_
    ↪ Encoder', 'Text Featurization Component.x'], 'Random Forest Classifier': ['Random_
    ↪ Forest Classifier', 'One Hot Encoder.x']}, 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)

```

(continues on next page)

(continued from previous page)

```

                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'],
↳ 'My Drop Null Columns Transformer': [MyDropNullColumns, 'Imputer.x'], 'DateTime_
↳ Featurization Component': ['DateTime Featurization Component', 'My Drop Null_
↳ Columns Transformer.x'], 'Text Featurization Component': ['Text Featurization_
↳ Component', 'DateTime Featurization Component.x'], 'One Hot Encoder': ['One Hot_
↳ Encoder', 'Text Featurization Component.x'], 'Random Forest Classifier': ['Random_
↳ Forest Classifier', 'One Hot Encoder.x']}, 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)

```

4.5 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.

4.5.1 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',
                                                                    test_size=0.
↳ 2, random_seed=0)

```

(continues on next page)

(continued from previous page)

```

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
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"})

      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)
```

(continued from previous page)

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'], 'Decision Tree Classifier': ['Decision Tree Classifier', 'Simple Imputer.
      ↪ x']}, 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)
```

(continues on next page)

(continued from previous page)

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]:

4.5.2 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

→ SHAP Value	Feature Name	Feature Value	Contribution to Prediction	→
→ 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	-	-

(continues on next page)

(continued from previous page)

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

	Feature Name	Feature Value	Contribution to Prediction
→ SHAP Value			
→	=====		
→ -0.04	worst perimeter	101.20	-
→ -0.05	worst concave points	0.06	-
→ -0.05	mean concave points	0.01	-

2 of 3

	Feature Name	Feature Value	Contribution to Prediction
→ SHAP Value			
→	=====		
→ 0.05	worst radius	11.94	-
→ 0.06	worst perimeter	80.78	-
→ 0.06	mean concave points	0.02	-

3 of 3

	Feature Name	Feature Value	Contribution to Prediction
→ SHAP Value			
→	=====		

(continues on next page)

(continued from previous page)

↪-0.05	worst concave points	0.10	-	↪
↪-0.06	worst perimeter	99.21	-	↪
↪-0.08	mean concave points	0.03	-	↪

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
```

↪SHAP Value	Feature Name	Feature Value	Contribution to Prediction	↪
↪=====				
↪-0.03	mean concavity	0.06	-	↪
↪-0.03	worst area	552.00	-	↪

(continues on next page)

(continued from previous page)

↪-0.05	worst concave points	0.08	-	↪
↪-0.05	worst radius	13.57	-	↪
↪-0.05	mean concave points	0.03	-	↪
↪-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				

(continues on next page)

(continued from previous page)

```

Predicted Probabilities: [benign: 1.0, malignant: 0.0]
Predicted Value: benign
Target Value: malignant
Cross Entropy: 7.987
Index ID: 135

```

	Feature Name	Feature Value	Contribution to Prediction
↪ SHAP Value			
↪	=====		
	mean concavity	0.05	-
↪ -0.03			
	worst area	653.60	-
↪ -0.04			
	worst concave points	0.09	-
↪ -0.05			
	worst radius	14.49	-
↪ -0.05			
	worst perimeter	92.04	-
↪ -0.06			
	mean concave points	0.03	-
↪ -0.06			

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
        }
      ]
    }
  ]
}
```

(continues on next page)

(continued from previous page)

```

    ]
  }
]
}

```

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",

```

(continues on next page)

(continued from previous page)

```

    "error_name": "Cross Entropy",
    "error_value": 0.0001970443507070075,
    "index_id": 475
  },
  "explanations": [
    {
      "feature_names": [
        "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",

```

(continues on next page)

(continued from previous page)

```

    "error_value": 7.986911819330411,
    "index_id": 135
  },
  "explanations": [
    {
      "feature_names": [
        "mean concavity",
        "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	-	
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

(continues on next page)

(continued from previous page)

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 0x7f4d50133400>
```

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 0x7f4d50062d90>
```

Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062100>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062b20>
Row: 1
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062790>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062e20>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062b50>
Row: 2
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062e80>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062f40>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d500626a0>
Row: 3
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062670>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062c70>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062400>
Row: 4
Class: class_0
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062340>
Class: class_1
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062fd0>
Class: class_2
<shap.plots._force.AdditiveForceVisualizer at 0x7f4d50062c10>

4.6 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.

4.6.1 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_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
```

4.6.2 Abnormal Data

EvalML provides a few data checks to check for abnormal data:

- `NoVarianceDataCheck`
- `ClassImbalanceDataCheck`
- `TargetLeakageDataCheck`
- `InvalidTargetDataCheck`
- `IDColumnsDataCheck`
- `OutliersDataCheck`
- `HighVarianceCVDataCheck`

- MulticollinearityDataCheck
- UniquenessDataCheck

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])
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])

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'])
```

(continues on next page)

(continued from previous page)

```
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'])
```

(continues on next page)

(continued from previous page)

```
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 Data Check

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,
                  '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 Data Check

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",
```

(continues on next page)

(continued from previous page)

```

                                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 Data Check

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.

```

4.6.3 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)

```

(continues on next page)

(continued from previous page)

```
# 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)

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.
```

4.6.4 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.

```
[13]: 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'])

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.
```

4.6.5 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.

```
[14]: 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])
```

4.6.6 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`
- `ClassImbalanceDataCheck` (for classification problem types)
- `NoVarianceDataCheck`
- `DateTimeNaNDataCheck`
- `NaturalLanguageNaNDataCheck`

4.6.7 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.

```
[15]: # Create a subclass of `DataChecks`
from evalml.data_checks import DataChecks, HighlyNullDataCheck,
↪InvalidTargetDataCheck, NoVarianceDataCheck, ClassImbalanceDataCheck,
↪TargetLeakageDataCheck
```

(continues on next page)

(continued from previous page)

```

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

```

```

[16]: # 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

```

4.7 Utilities

4.7.1 Configuring Logging

EvalML uses [the standard python logging package](#). By default, EvalML will log INFO-level logs and higher (warnings, errors and critical) to stdout, and will log everything to `evalml_debug.log` in the current working directory.

If you want to change the location of the logfile, before import, set the `EVALML_LOG_FILE` environment variable to specify a filename within an existing directory in which you have write permission. If you want to disable logging to the logfile, set `EVALML_LOG_FILE` to be empty. If the environment variable is set to an invalid location, EvalML will print a warning message to stdout and will not create a log file.

4.7.2 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: 0.29.0
EvalML installation directory: /home/docs/checkouts/readthedocs.org/user_builds/
    ↪feature-labs-inc-evalml/envs/v0.29.0/lib/python3.8/site-packages/evalml

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: 6.1G

INSTALLED VERSIONS
-----
zict: 2.0.0
xgboost: 1.4.2
wrapt: 1.12.1
woodwork: 0.5.0
widgetsnbextension: 3.5.1
wheel: 0.36.2
webencodings: 0.5.1
wcwidth: 0.2.5
urllib3: 1.26.6
unicodecode: 1.2.0
traitlets: 5.0.5
tqdm: 4.61.2
tornado: 6.1
toolz: 0.11.1
threadpoolctl: 2.2.0
```

(continues on next page)

(continued from previous page)

```
texttable: 1.6.4
testpath: 0.5.0
terminado: 0.10.1
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: 1.8.1
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: 57.4.0
send2trash: 1.7.1
seaborn: 0.11.1
scipy: 1.7.0
scikit-optimize: 0.8.1
scikit-learn: 0.24.2
requirements-parser: 0.2.0
requests: 2.26.0
regex: 2021.7.6
recommonmark: 0.5.0
readthedocs-sphinx-ext: 2.1.4
pyzmq: 22.1.0
pyyaml: 5.4.1
pytz: 2021.1
python-dateutil: 2.8.2
pysistent: 0.18.0
pyparsing: 2.4.7
pygments: 2.9.0
pydata-sphinx-theme: 0.6.3
pyparser: 2.20
pyaml: 20.4.0
ptyprocess: 0.7.0
psutil: 5.8.0
prompt-toolkit: 3.0.19
prometheus-client: 0.11.0
pmdarima: 1.8.0
plotly: 5.1.0
pip: 21.1.3
pillow: 8.3.1
pickleshare: 0.7.5
pexpect: 4.8.0
patsy: 0.5.1
partd: 1.2.0
parso: 0.8.2
```

(continues on next page)

(continued from previous page)

```
pandocfilters: 1.4.3
pandas: 1.3.0
packaging: 21.0
numpy: 1.21.1
numba: 0.53.1
notebook: 6.4.0
nltk: 3.6.2
nlp-primitives: 1.1.0
networkx: 2.5.1
nest-asyncio: 1.5.1
nbsphinx: 0.8.6
nbformat: 5.1.3
nbconvert: 6.1.0
nbclient: 0.5.3
msgpack: 1.0.2
mock: 1.0.1
mistune: 0.8.4
matplotlib: 3.4.2
matplotlib-inline: 0.1.2
markupsafe: 2.0.1
loket: 0.2.1
llvmlite: 0.36.0
lightgbm: 3.2.1
lazy-object-proxy: 1.6.0
kiwisolver: 1.3.1
kaleido: 0.2.1
jupyterlab-widgets: 1.0.0
jupyterlab-pygments: 0.1.2
jupyter-core: 4.7.1
jupyter-client: 6.1.12
jsonschema: 3.2.0
joblib: 1.0.1
jinja2: 3.0.1
jedi: 0.18.0
ipywidgets: 7.6.3
ipython: 7.25.0
ipython-genutils: 0.2.0
ipykernel: 6.0.3
imbalanced-learn: 0.8.0
imagesize: 1.2.0
idna: 3.2
heapdict: 1.0.1
graphviz: 0.17
future: 0.18.2
fsspec: 2021.7.0
featuretools: 0.26.0
evalml: 0.29.0
entrypoints: 0.3
docutils: 0.16
distributed: 2021.7.0
defusedxml: 0.7.1
decorator: 4.4.2
debugpy: 1.4.0
dask: 2021.7.0
cython: 0.29.17
cycller: 0.10.0
commonmark: 0.8.1
```

(continues on next page)

(continued from previous page)

```
colorama: 0.4.4
cloudpickle: 1.6.0
click: 8.0.1
charset-normalizer: 2.0.3
cffi: 1.14.6
certifi: 2021.5.30
category-encoders: 2.2.2
catboost: 0.26
bleach: 3.3.1
beautifulsoup4: 4.9.3
backcall: 0.2.0
babel: 2.9.1
attrs: 21.2.0
async-generator: 1.10
astroid: 2.6.5
argon2-cffi: 20.1.0
alabaster: 0.7.12
```

4.8 FAQ

4.8.1 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.

4.8.2 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`).

4.8.3 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.

4.8.4 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`).

4.8.5 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).

4.8.6 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.

4.8.7 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.

4.8.8 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.

4.8.9 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.

4.8.10 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>drop_nan_target_rows</i>	Drops rows in X and y when row in the target y has a value of NaN.
<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 is 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 all_components()

continues on next page

Table 3 – continued from previous page

<i>ComponentNotYetFittedError</i>	An exception to be raised when predict/predict_proba/transform is called on a component without fitting first.
<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

<i>silent_error_callback</i>	No-op.
<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.

5.5 Pipelines

5.5.1 Pipeline Base Classes

<i>PipelineBase</i>	Machine learning pipeline made out of transformers and a estimator.
<i>ClassificationPipeline</i>	Pipeline subclass for all classification pipelines.
<i>BinaryClassificationPipeline</i>	Pipeline subclass for all binary classification pipelines.
<i>MulticlassClassificationPipeline</i>	Pipeline subclass for all multiclass classification pipelines.
<i>RegressionPipeline</i>	Pipeline subclass for all regression pipelines.
<i>TimeSeriesClassificationPipeline</i>	Pipeline base class for time series classification problems.
<i>TimeSeriesBinaryClassificationPipeline</i>	Pipeline base class for time series binary 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.

5.5.2 Pipeline Utils

<i>make_pipeline</i>	Given input data, target data, an estimator class and the problem type,
<i>generate_pipeline_code</i>	Creates and returns a string that contains the Python imports and code required for running the EvalML pipeline.

5.6 Components

5.6.1 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.

continues on next page

Table 10 – continued from previous page

<i>Estimator</i>	A component that fits and predicts given data.
------------------	------------------------------------------------

5.6.2 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.

5.6.3 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>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.
<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.

continues on next page

Table 12 – continued from previous page

<i>SMOTESampler</i>	SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.
<i>SMOTENCSampler</i>	SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns.
<i>SMOTENSampler</i>	SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets.

5.6.4 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>StackedEnsembleClassifier</i>	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>StackedEnsembleRegressor</i>	Stacked Ensemble Regressor.

continues on next page

Table 14 – continued from previous page

<i>DecisionTreeRegressor</i>	Decision Tree Regressor.
<i>LightGBMRegressor</i>	LightGBM Regressor.
<i>SVMRegressor</i>	Support Vector Machine Regressor.

5.7 Model Understanding

5.7.1 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 <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> .
<i>get_linear_coefficients</i>	Returns a dataframe showing the features with the greatest predictive power for a linear model.
<i>t_sne</i>	Get the transformed output after fitting X to the embedded space using t-SNE.

5.7.2 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.

continues on next page

Table 16 – continued from previous page

<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 .

5.7.3 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.

5.8 Objectives

5.8.1 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.

5.8.2 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

5.8.3 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.

continues on next page

Table 20 – continued from previous page

<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>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.

5.8.4 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.

5.8.5 Objective Utils

<code>get_all_objective_names</code>	Get a list of the names of all objectives.
<code>get_core_objectives</code>	Returns all core objective instances associated with the given problem type.
<code>get_core_objective_names</code>	Get a list of all valid core objectives.
<code>get_non_core_objectives</code>	Get non-core objective classes.
<code>get_objective</code>	Returns the Objective class corresponding to a given objective name.

5.9 Problem Types

<code>handle_problem_types</code>	Handles <code>problem_type</code> by either returning the <code>ProblemTypes</code> or converting from a str.
<code>detect_problem_type</code>	Determine the type of problem is being solved based on the targets (binary vs multiclass classification, regression)
<code>ProblemTypes</code>	Enum defining the supported types of machine learning problems.

5.10 Model Family

<code>handle_model_family</code>	Handles <code>model_family</code> by either returning the <code>ModelFamily</code> or converting from a string
<code>ModelFamily</code>	Enum for family of machine learning models.

5.11 Tuners

<code>Tuner</code>	Defines API for base Tuner classes.
<code>SKOptTuner</code>	Bayesian Optimizer.
<code>GridSearchTuner</code>	Grid Search Optimizer, which generates all of the possible points to search for using a grid.
<code>RandomSearchTuner</code>	Random Search Optimizer.

5.12 Data Checks

5.12.1 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>DataChecks</i>	A collection of data checks.
<i>DefaultDataChecks</i>	A collection of basic data checks that is used by AutoML by default.

5.12.2 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.

5.12.3 Data Check Message Types

<i>DataCheckMessageType</i>	Enum for type of data check message: WARNING or ERROR.
-----------------------------	--------------------------------------------------------

5.12.4 Data Check Message Codes

<i>DataCheckMessageCode</i>	Enum for data check message code.
-----------------------------	-----------------------------------

5.13 Utils

5.13.1 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, 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

<i>add_result</i>	Register results from evaluating a pipeline
-------------------	---------------------------------------------

continues on next page

Table 33 – continued from previous page

<i>batch_number</i>	Returns the number of batches which have been recommended so far.
<i>next_batch</i>	Get the next batch of pipelines to evaluate
<i>pipeline_number</i>	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

iterative_algorithm

Module Contents

Classes Summary

<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.
---------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------

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_size=None,
max_ensemble_size=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>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.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

<code>CFClient</code>	Custom CFClient API to match Dask's CFClient and allow context management.
-----------------------	----------------------------------------------------------------------------

continues on next page

Table 40 – continued from previous page

<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>submit</i>	Pass through to imitate Dask’s Client API.
---------------	--------------------------------------------

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.
<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)`

The concurrent.futures (CF) engine

Methods

`setup_job_log`

`submit_evaluation_job` Send evaluation job to cluster.

`submit_scoring_job` Send scoring job to cluster.

`submit_training_job` Send training job to cluster.

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** (*client*)

The dask engine

Methods

<i>send_data_to_cluster</i>	Send data to the cluster.
<i>setup_job_log</i>	
<i>submit_evaluation_job</i>	Send evaluation job to cluster.
<i>submit_scoring_job</i>	Send scoring job to cluster.
<i>submit_training_job</i>	Send training job to cluster.

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

<i>evaluate_pipeline</i>	Function submitted to the <code>submit_evaluation_job</code> engine method.
<i>score_pipeline</i>	Wrapper around <code>pipeline.score</code> method to make it easy to score pipelines with dask.
<i>train_and_score_pipeline</i>	Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores
<i>train_pipeline</i>	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

<i>setup_job_log</i>	
<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.

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

<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 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

<i>debug</i>	Store message at the debug level.
<i>error</i>	Store message at the error level.
<i>info</i>	Store message at the info level.
<i>warning</i>	Store message at the warning level.
<i>write_to_logger</i>	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

<i>cancel</i>	Cancel the current computation.
<i>done</i>	Whether the computation is done.
<i>get_result</i>	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

<i>setup_job_log</i>	
<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.

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

<i>CFEngine</i>	The concurrent.futures (CF) engine
<i>DaskEngine</i>	The dask engine
<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>SequentialEngine</i>	The default engine for the AutoML search. Trains and scores pipelines locally and sequentially.

Functions

<i>evaluate_pipeline</i>	Function submitted to the submit_evaluation_job engine method.
<i>train_and_score_pipeline</i>	Given a pipeline, config and data, train and score the pipeline and return the CV or TV scores
<i>train_pipeline</i>	Train a pipeline and tune the threshold if necessary.

Contents

class evalml.automl.engine.CFEngine (*client*)

The concurrent.futures (CF) engine

Methods

<i>setup_job_log</i>	
<i>submit_evaluation_job</i>	Send evaluation job to cluster.
<i>submit_scoring_job</i>	Send scoring job to cluster.
<i>submit_training_job</i>	Send training job to cluster.

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* (*client*)

The dask engine

Methods

<i>send_data_to_cluster</i>	Send data to the cluster.
<i>setup_job_log</i>	
<i>submit_evaluation_job</i>	Send evaluation job to cluster.
<i>submit_scoring_job</i>	Send scoring job to cluster.
<i>submit_training_job</i>	Send training job to cluster.

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

<i>setup_job_log</i>	
<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.

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

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*()

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

<code>AutoMLSearch</code>	Automated Pipeline search.
---------------------------	----------------------------

Functions

<code>search</code>	Given data and configuration, run an automl search.
---------------------	-----------------------------------------------------

Attributes Summary

<code>logger</code>

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, engine=None)
```

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`, 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.
- **engine** (*EngineBase* or *None*) – The engine instance used to evaluate pipelines. If `None`, a `SequentialEngine` will be used.

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 <code>False</code> , returns an untrained pipeline instance.
<code>describe_pipeline</code>	Describe a pipeline
<code>full_rankings</code>	Returns a <code>pandas.DataFrame</code> 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>	

continues on next page

Table 65 – continued from previous page

<i>rankings</i>	Returns a pandas.DataFrame with scoring results from the highest-scoring set of parameters used with each pipeline.
<i>results</i>	Class that allows access to a copy of the results from <i>automl_search</i> .
<i>save</i>	Saves AutoML object at file path
<i>score_pipelines</i>	Score a list of pipelines on the given holdout data.
<i>search</i>	Find the best pipeline for the data set.
<i>train_pipelines</i>	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

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 logger`

`evalml automl automl_search search` (*X_train=None, y_train=None, problem_type=None, objective='auto', **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.

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

`log_error_callback`

Logs the exception thrown as an error. Will not throw. This is the default behavior for AutoMLSearch.

continues on next page

Table 66 – continued from previous page

<code>raise_error_callback</code>	Raises the exception thrown by the AutoMLSearch object. Also logs the exception as an error.
<code>silent_error_callback</code>	No-op.

Attributes Summary

<code>logger</code>

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.pipeline_search_plots.**SearchIterationPlot** (*results*, *objective*)

Methods

update

update (*self*, *results*, *objective*)

utils

Module Contents

Functions

<i>check_all_pipeline_names_unique</i>	Checks whether all the pipeline names are unique.
<i>get_best_sampler_for_data</i>	Returns the name of the sampler component to use for AutoMLSearch.
<i>get_default_primary_search_objective</i>	Get the default primary search objective for a problem type.
<i>get_hyperparameter_ranges</i>	Returns hyperparameter ranges from all components as a dictionary.
<i>get_pipelines_from_component_graphs</i>	Returns created pipelines from passed component graphs based on the specified 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>tune_binary_threshold</i>	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_hyperparameter_ranges (component_graph, custom_hyperparameters)
```

Returns hyperparameter ranges from all components as a dictionary.

Parameters

- **component_graph** (*list(str, ComponentBase)*) – The component_graph of the pipeline.
- **custom_hyperparameters** (*dict*) – The custom hyperparameters to be passed to the pipeline.

Returns Dictionary of hyperparameter ranges for each component in the component graph.

Return type dict

```
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

<code>get_default_primary_search_objective</code>	Get the default primary search objective for a problem type.
<code>make_data_splitter</code>	Given the training data and ML problem parameters, compute a data splitting method to use during AutoML search.
<code>search</code>	Given data and configuration, run an automl search.
<code>tune_binary_threshold</code>	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,
                                engine=None)
```

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`, 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.
- **engine** (*EngineBase* or *None*) – The engine instance used to evaluate pipelines. If None, a `SequentialEngine` will be used.

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>describe_pipeline</code>	Describe a pipeline

continues on next page

Table 75 – continued from previous page

<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*

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

<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.

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', **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.

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

`setup_job_log`

`submit_evaluation_job`

Submit job for pipeline evaluation during AutoMLSearch.

continues on next page

Table 77 – continued from previous page

<code>submit_scoring_job</code>	Submit job for pipeline scoring.
<code>submit_training_job</code>	Submit job for pipeline training.

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 \bar{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

<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": []}]}
```

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.
-------------	-------------------------------------------

continues on next page

Table 81 – continued from previous page

<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.

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

message_type	DataCheckMessageType.ERROR
---------------------	----------------------------

Methods

to_dict

to_dict (*self*)

```
class evalml.data_checks.data_check_message.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.data_check_message.DataCheckWarning (message,  
data_check_name,  
message_code=None,  
details=None)
```

DataCheckMessage subclass for warnings returned by data checks.

Attributes

message_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	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.
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 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_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 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_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_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()
>>> 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)

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*

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.

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_threshold=0.95*)
Checks if there are any highly-null columns and rows in the input.

Parameters **pct_null_threshold** (*float*) – If the percentage of NaN values in an input feature exceeds this amount, that column/row will be considered highly-null. Defaults to 0.95.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	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_threshold=0.50)
```

(continues on next page)

(continued from previous page)

```

>>> 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"}}}]

```

id_columns_data_check

Module Contents

Classes Summary

<i>IDColumnsDataCheck</i>	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

<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"}},
↳         "actions": [{"code": "DROP_COL",
↳                       "metadata": {"column": "df_id
↳ }}}}
↳ }]]}
```

invalid_targets_data_check

Module Contents

Classes Summary

<i>InvalidTargetDataCheck</i>	Checks if the target data contains missing or invalid values.
-------------------------------	---------------------------------------------------------------

Contents

class evalml.data_checks.invalid_targets_data_check.**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",
↳                                     "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

<i>MulticollinearityDataCheck</i>	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, np.ndarray*) – The input features to check

Returns dict with a DataCheckWarning if there are any potentially multicollinear columns.

Return type dict

natural_language_nan_data_check

Module Contents

Classes Summary

<i>NaturalLanguageNaNDataCheck</i>	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": [],
...     "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,
```

(continues on next page)

(continued from previous page)

```
...         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.
----------------------------	--------------------------------------------------------------

Attributes Summary

<i>logger</i>	
---------------	--

Contents

evalml.data_checks.no_variance_data_check.logger

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",
↪                                     "code": "HAS_OUTLIERS",
↪                                     "details": {"columns": ["z"]}}]},
↪                                     "actions": []}
```

(continues on next page)

(continued from previous page)

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":
↳                                     "code
↳"warning",
↳": "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_leakage_data_check

Module Contents

Classes Summary

TargetLeakageDataCheck

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": [],
```

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",
↳                                     "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 {} problem type)
```

```
evalml.data_checks.uniqueness_data_check.warning_too_unique = Input columns ({} for {} problem type)
```

utils

Module Contents

Classes Summary

EmptyDataChecks

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>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.

continues on next page

Table 127 – continued from previous page

<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>TargetLeakageDataCheck</i>	Check if any of the features are highly correlated with the target by using mutual information or Pearson correlation.
<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])
>>> 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.
-------------	-------------------------------------------

continues on next page

Table 129 – continued from previous page

<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.
IMPUTE_COL	Action code for imputing a column.

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

message_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.
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 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_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 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_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

`to_dict`

`to_dict (self)`
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

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	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*)

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*

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.

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_threshold=0.95*)

Checks if there are any highly-null columns and rows in the input.

Parameters **pct_null_threshold** (*float*) – If the percentage of NaN values in an input feature exceeds this amount, that column/row will be considered highly-null. Defaults to 0.95.

Methods

<i>name</i>	Returns a name describing the data check.
<i>validate</i>	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_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":
```

(continues on next page)

(continued from previous page)

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"}},
↳ "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",
↳                                     "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

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	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*, *np.ndarray*) – The input features to check

Returns dict with a DataCheckWarning if there are any potentially multicollinear columns.

Return type dict

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

<code>name</code>	Returns a name describing the data check.
<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()
```

(continues on next page)

(continued from previous page)

```
>>> 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

<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

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"]}]},
↳                                     "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.**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

<code>name</code>	Returns a name describing the data check.
<code>validate</code>	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

<code>load_breast_cancer</code>	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)

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_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_wine()
```

Load wine dataset. Multiclass problem.

Returns X and y

Return type (pd.DataFrame, pd.Series)

Exceptions

Submodules

exceptions

Module Contents

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 is 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.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

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 is 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.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

ModelFamily

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
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.

utils

Module Contents

Functions

<code>handle_model_family</code>	Handles <code>model_family</code> by either returning the <code>ModelFamily</code> 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

<code>ModelFamily</code>	Enum for family of machine learning models.
--------------------------	---------------------------------------------

Functions

<code>handle_model_family</code>	Handles <code>model_family</code> by either returning the <code>ModelFamily</code> 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
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=
out-
put_format=)
```

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”.

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.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

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	compute_feature_stage
COM- PUTE_SHAP_VALUES_STAGE	compute_shap_value_stage
DONE	done
PRE- DICT_STAGE	predict_stage
PREPRO- CESS- ING_STAGE	preprocessing_stage

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')
```

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”.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features to each prediction for each row of the input data. 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="text",
callback=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

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': AdditiveForceVisualizer},
 {'malignant': {'expected_value': 0.29, 'feature_names': ['worst concave points',
                                                         'worst perimeter', 'worst radius'], 'shap_values': [0.05, 0.03, 0.02], 'plot': AdditiveForceVisualizer}}
```

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*, *matplotlib=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 <i>y_true</i> and <i>y_pred</i> into a single dataframe and adds a column for outliers. Used in <code>graph_prediction_vs_actual()</code> .
<code>get_prediction_vs_actual_over_time_data</code>	Get the data needed for the <code>prediction_vs_actual_over_time</code> 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

continues on next page

Table 169 – continued from previous page

<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,
                                                                           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.
- **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, nor-
                                                         malize_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 ValueError – if a graph is requested for a class name that isn’t present in the pipeline

```
evalml.model_understanding.graphs.graph_permutation_importance(pipeline, X,  
                                                                y, objective,  
                                                                importance,  
                                                                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,
                                                                           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.
- **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, 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.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, per-
centiles=(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

- **ValueError** – if the user provides a tuple of not exactly two features.
- **ValueError** – if the provided pipeline isn’t fitted.
- **ValueError** – if the provided pipeline is a Baseline pipeline.
- **ValueError** – if any of the features passed in are completely NaN
- **ValueError** – 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=None, n_jobs=None, random_seed=None*)

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_col</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.

continues on next page

Table 171 – continued from previous page

<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.
<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> .
<code>get_prediction_vs_actual_over_time_data</code>	Get the data needed for the <code>prediction_vs_actual_over_time</code> 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.

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, objective, n_repeats=5, fast=True, precomputed_features=None, random_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, 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.explain_predictions(pipeline, input_features, y, indices_to_explain, top_k_features=3, include_shap_values=False, include_expected_value=False, output_format='text')
```

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”.

Returns

str, dict, or pd.DataFrame - A report explaining the top contributing features to each prediction for each row of the dataset. 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)
```

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

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.

`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, 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.
- **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 **ValueError** – if a graph is requested for a class name that isn’t present in the pipeline

`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, 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.
- **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, 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.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

- **ValueError** – if the user provides a tuple of not exactly two features.
- **ValueError** – if the provided pipeline isn't fitted.
- **ValueError** – if the provided pipeline is a Baseline pipeline.
- **ValueError** – if any of the features passed in are completely NaN
- **ValueError** – 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>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.
<i>score</i>	Returns a numerical score indicating performance based on the differences between the predicted and actual values.

continues on next page

Table 173 – continued from previous page

<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 `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

<i>CostBenefitMatrix</i>	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

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	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>	Calculates cost-benefit of the using the predicted and true values.

continues on next page

Table 175 – 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*)

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

FraudCost

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

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.
<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

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

greater_is_better	True
is_bounded_like_percentage	False
name	Lead Scoring
perfect_score	None
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>	Calculate the profit per lead.
<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*)

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

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
<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.

continues on next page

Table 181 – continued from previous page

<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 `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>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 183 – 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 `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

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>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 `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

<i>SensitivityLowAlert</i>	Base class for all binary classification objectives.
----------------------------	------------------------------------------------------

Attributes Summary

<i>logger</i>

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

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	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>	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.

continues on next page

Table 188 – 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

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.

continues on next page

Table 189 – continued from previous page

<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>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>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

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

<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.**AccuracyMulticlass**

Accuracy score for multiclass classification.

Attributes

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

<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.**AUC**

AUC score for binary classification.

Attributes

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

continues on next page

Table 192 – 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.standard_metrics.**AUCMacro**

AUC score for multiclass classification using macro averaging.

Attributes

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

<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.**AUCMicro**

AUC score for multiclass classification using micro averaging.

Attributes

greater_is_better	True
is_bounded_like_percentage	True
name	AUC 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.**AUCWeighted**

AUC Score for multiclass classification using weighted averaging.

Attributes

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

<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.**BalancedAccuracyBinary**

Balanced accuracy score for binary classification.

Attributes

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

continues on next page

Table 196 – 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.standard_metrics.**BalancedAccuracyMulticlass**

Balanced accuracy score for multiclass classification.

Attributes

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.standard_metrics.**ExpVariance**

Explained variance score for regression.

Attributes

greater_is_better	True
is_bounded_like_percentage	False
name	ExpVariance
perfect_score	1.0
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.standard_metrics.F1

F1 score for binary classification.

Attributes

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

<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

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

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

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	

continues on next page

Table 201 – 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.standard_metrics.**F1Weighted**

F1 score for multiclass classification using weighted averaging.

Attributes

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

<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.**LogLossBinary**

Log Loss for binary classification.

Attributes

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

<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.**LogLossMulticlass**

Log Loss for multiclass classification.

Attributes

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.standard_metrics.**MAE**

Mean absolute error for regression.

Attributes

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.

continues on next page

Table 205 – 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.**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

greater_is_better	False
is_bounded_like_percentage	True
name	Mean Absolute Percentage Error
perfect_score	0.0
problem_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
<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.standard_metrics.**MaxError**

Maximum residual error for regression.

Attributes

greater_is_better	False
is_bounded_like_percentage	False
name	MaxError
perfect_score	0.0
problem_types	[ProblemTypes.REGRESSION, ProblemTypes.TIME_SERIES_REGRESSION]
score_needs_problem	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.**MCCBinary**

Matthews correlation coefficient for binary classification.

Attributes

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	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.**MCCMulticlass**

Matthews correlation coefficient for multiclass classification.

Attributes

greater_is_better	True
is_bounded_like_percentage	False
name	MCC Multiclass
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.**MeanSquaredLogError**

Mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError

Attributes

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_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
<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.standard_metrics.**MedianAE**

Median absolute error for regression.

Attributes

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_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 211 – 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.standard_metrics.**MSE**

Mean squared error for regression.

Attributes

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.standard_metrics.**Precision**

Precision score for binary classification.

Attributes

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	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.**PrecisionMacro**

Precision score for multiclass classification using macro averaging.

Attributes

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.standard_metrics.**PrecisionMicro**

Precision score for multiclass classification using micro averaging.

Attributes

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	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 215 – 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.**PrecisionWeighted**

Precision score for multiclass classification using weighted averaging.

Attributes

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	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.R2`

Coefficient of determination for regression.

Attributes

greater_is_better	True
is_bounded_like_percentage	False
name	R2
perfect_score	1
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>	

continues on next page

Table 217 – 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.standard_metrics.**Recall**

Recall score for binary classification.

Attributes

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

<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

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	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.**RecallMicro**

Recall score for multiclass classification using micro averaging.

Attributes

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

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	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 221 – 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

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_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

greater_is_better	False
is_bounded_like_percentage	False
name	Root Mean Squared Log Error
perfect_score	0.0
problem_types	[<code>ProblemTypes.REGRESSION</code> , <code>ProblemTypes.TIME_SERIES_REGRESSION</code>]
score_needs_positive	True

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
-----------------------------------------------------------	--------------------------------------------------

continues on next page

Table 223 – 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

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>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.

continues on next page

Table 225 – 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 `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

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.
<i>get_objective</i>	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

<code>AccuracyBinary</code>	Accuracy score for binary classification.
<code>AccuracyMulticlass</code>	Accuracy score for multiclass classification.
<code>AUC</code>	AUC score for binary classification.

continues on next page

Table 227 – continued from previous page

<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>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.
<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.

continues on next page

Table 227 – continued from previous page

<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

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>	
<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.

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

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

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

<i>calculate_percent_difference</i>	Calculate the percent difference between scores.
-----------------------------------------------------	--------------------------------------------------

continues on next page

Table 230 – 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 (*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

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.**AUCMacro**

AUC score for multiclass classification using macro averaging.

Attributes

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
<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.AUCMicro

AUC score for multiclass classification using micro averaging.

Attributes

greater_is_better	True
is_bounded_like_percentage	True
name	AUC 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.**AUCWeighted**

AUC Score for multiclass classification using weighted averaging.

Attributes

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.

continues on next page

Table 234 – 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.BalancedAccuracyBinary

Balanced accuracy score for binary classification.

Attributes

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.**BalancedAccuracyMulticlass**

Balanced accuracy score for multiclass classification.

Attributes

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

<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.**BinaryClassificationObjective**

Base class for all binary classification objectives.

Attributes

problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
----------------------	--------------------------------------------------------

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>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>optimize_threshold</code>	Learn a binary classification threshold which optimizes the current objective.
<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 `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 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

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	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 238 – 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>	Calculates cost-benefit of the using the predicted and true values.
<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*)

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

greater_is_better	True
is_bounded_like_percentage	False
name	ExpVariance
perfect_score	1.0
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.**F1**

F1 score for binary classification.

Attributes

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

<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

continues on next page

Table 240 – 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.**F1Macro**

F1 score for multiclass classification using macro averaging.

Attributes

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

<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.F1Micro

F1 score for multiclass classification using micro averaging.

Attributes

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.
<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

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

<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.**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

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

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.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

greater_is_better	True
is_bounded_like_percentage	False
name	Lead Scoring
perfect_score	None
problem_types	[ProblemTypes.BINARY, ProblemTypes.TIME_SERIES_BINARY]
score_needs_positive	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>	Calculate the profit per lead.
<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*)
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

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

<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.**LogLossMulticlass**

Log Loss for multiclass classification.

Attributes

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.MAE`

Mean absolute error for regression.

Attributes

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

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<code>is_defined_for_problem_type</code>	

continues on next page

Table 248 – 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.**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

greater_is_better	False
is_bounded_like_percentage	True
name	Mean Absolute Percentage Error
perfect_score	0.0
problem_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
<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

greater_is_better	False
is_bounded_like_percentage	False
name	MaxError
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 (*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

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

<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

greater_is_better	True
is_bounded_like_percentage	False
name	MCC Multiclass
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 252 – 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.**MeanSquaredLogError**

Mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError

Attributes

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_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
<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

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_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.**MSE**

Mean squared error for regression.

Attributes

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

problem_types	[ProblemTypes.MULTICLASS, ProblemTypes.TIME_SERIES_MULTICLASS]
----------------------	----------------------------------------------------------------

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<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 `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

prob- lem_types	None
----------------------------	------

Methods

<code>calculate_percent_difference</code>	Calculate the percent difference between scores.
<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 `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.Precision

Precision score for binary classification.

Attributes

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

<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.PrecisionMacro

Precision score for multiclass classification using macro averaging.

Attributes

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	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.PrecisionMicro`

Precision score for multiclass classification using micro averaging.

Attributes

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	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 260 – 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.PrecisionWeighted

Precision score for multiclass classification using weighted averaging.

Attributes

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	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.**R2**

Coefficient of determination for regression.

Attributes

greater_is_better	True
is_bounded_like_percentage	False
name	R2
perfect_score	1
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 (*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

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

<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.**RecallMacro**

Recall score for multiclass classification using macro averaging.

Attributes

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	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.RecallMicro

Recall score for multiclass classification using micro averaging.

Attributes

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.RecallWeighted

Recall score for multiclass classification using weighted averaging.

Attributes

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	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.

continues on next page

Table 266 – 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.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>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 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

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_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.**RootMeanSquaredLogError**

Root mean squared log error for regression.

Only valid for nonnegative inputs. Otherwise, will throw a ValueError.

Attributes

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

<i>calculate_percent_difference</i>	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.**SensitivityLowAlert** (*alert_rate=0.01*)

Base class for all binary classification objectives.

Attributes

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

stacked_ensemble_base

Module Contents

Classes Summary

StackedEnsembleBase

Stacked Ensemble Base Class.

Contents

```
class evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase (input_p
fi-
nal_esti
cv=None
n_jobs=

1,
ran-
dom_see
**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.
--------------	-----------------------------------------------------------------------

continues on next page

Table 272 – continued from previous page

<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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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

stacked_ensemble_classifier

Module Contents

Classes Summary

StackedEnsembleClassifier

Stacked Ensemble Classifier.

Contents

class evalml.pipelines.components.ensemble.stacked_ensemble_classifier.**StackedEnsembleClass**

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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

stacked_ensemble_regressor

Module Contents

Classes Summary

StackedEnsembleRegressor

Stacked Ensemble Regressor.

Contents

class evalml.pipelines.components.ensemble.stacked_ensemble_regressor.**StackedEnsembleRegressor**

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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>StackedEnsembleBase</i>	Stacked Ensemble Base Class.
<i>StackedEnsembleClassifier</i>	Stacked Ensemble Classifier.
<i>StackedEnsembleRegressor</i>	Stacked Ensemble Regressor.

Contents

```
class evalml.pipelines.components.ensemble.StackedEnsembleBase (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

continues on next page

Table 278 – continued from previous page

<i>feature_importance</i>	Not implemented for StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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.StackedEnsembleClassifier (input_pipelines=None,
                                                                    fi-
                                                                    nal_estimator=None,
                                                                    cv=None,
                                                                    n_jobs=-
                                                                    1, ran-
                                                                    dom_seed=0,
                                                                    **kwargs)
```

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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.ensemble.StackedEnsembleRegressor (input_pipelines=None,
                                                                    final_estimator=None,
                                                                    cv=None,
                                                                    n_jobs=-1,
                                                                    random_seed=0,
                                                                    **kwargs)
```

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	Stacked Ensemble 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 stacked ensemble classes.
<i>describe</i>	Describe a component and its parameters
<i>feature_importance</i>	Not implemented for StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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

<i>BaselineClassifier</i>	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
predict_uses_y	False
supported_problem_types	[ProblemTypes.BINARY, ProblemTypes.MULTICLASS]

Methods

<code>classes_</code>	Returns class labels. Will return None before fitting.
<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. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<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

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

<i>CatBoostClassifier</i>	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
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

elasticnet_classifier

Module Contents

Classes Summary

<i>ElasticNetClassifier</i>	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

<i>ExtraTreesClassifier</i>	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

<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`

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

<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

continues on next page

Table 292 – 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 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 $\text{bagging_fraction} * 100\%$ 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 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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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 300 – 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 -1.

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.

continues on next page

Table 302 – 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.

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', 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
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=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
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_split=  
                                                                    min_weight_fraction=  
                                                                    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.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,
**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
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,  
                                                                    kernel='rbf',  
                                                                    gamma='scale',  
                                                                    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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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": ["linear", "poly", "rbf", "sigmoid", "precomputed"], "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,
                                                                    n_jobs=-1,
                                                                    **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 -1.

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 314 – 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

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_index, d=0, start_season, max_p, max_d, max_q, seasonal, n_jobs, random_seed*)

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 318 – 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*, 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

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

rf_regressor

Module Contents

Classes Summary

<i>RandomForestRegressor</i>	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

<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 332 – continued from previous page

<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

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='scale',
                                                                    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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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 334 – 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

TimeSeriesBaselineEstimator

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.
- **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
pre-dict_uses_y	True
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*, *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*, *y=None*)

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

<i>XGBoostRegressor</i>	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 -1.

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>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',
                                                                    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
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. Since baseline regressors 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

continues on next page

Table 341 – 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. 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=0.01,
                                                                    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

<i>clone</i>	Constructs a new component with the same parameters and random state.
--------------	-----------------------------------------------------------------------

continues on next page

Table 344 – 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 345 – 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 346 – 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.RandomForestRegressor (n_estimators=100
                                                                              max_depth=6,
                                                                              n_jobs=-
                                                                              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
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.SVMRegressor (C=1.0,
                                                                    kernel='rbf',
                                                                    gamma='scale',
                                                                    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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses 1 / (n_features * X.var()) as value of gamma - if “auto”, 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": ["linear", "poly", "rbf", "sigmoid", "precomputed"], "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.regressors.TimeSeriesBaselineEstimator (gap=1,  

random_seed=None,  

**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.
- **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	True
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*, *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*, *y=None*)

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,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    n_jobs=-  
                                                                    1,  
                                                                    **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 -1.

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

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.

continues on next page

Table 353 – 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>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

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.

continues on next page

Table 354 – continued from previous page

<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>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,      sea-
                                                             sonal=True,
                                                             n_jobs=- 1,   ran-
                                                             dom_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.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

<code>classes_</code>	Returns class labels. Will return <code>None</code> before fitting.
<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. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<code>fit</code>	Fits component to data
<code>load</code>	Loads component at file path

continues on next page

Table 356 – 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

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
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. Since baseline regressors 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

clone (*self*)

Constructs a new component with the same parameters 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*, 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.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
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=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
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.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.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 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

```
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

<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

continues on next page

Table 363 – 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 366 – 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-paramer-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 370 – 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_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.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

continues on next page

Table 372 – 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.RandomForestRegressor (n_estimators=100,  
                                                                max_depth=6,  
                                                                n_jobs=-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.SVMClassifier(C=1.0, kernel='rbf',
                                                         gamma='scale',
                                                         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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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

<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 374 – 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.SVMRegressor (C=1.0, kernel='rbf',
                                                         gamma='scale',
                                                         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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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

<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.**TimeSeriesBaselineEstimator** (*gap=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.
- **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
pre-dict_uses_y	True
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*, *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*, *y=None*)

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,
                                                             random_seed=0,
                                                             n_jobs=-1,
                                                             **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 -1.

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

continues on next page

Table 377 – 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.XGBoostRegressor (eta=0.1,  
                                                             max_depth=6,  
                                                             min_child_weight=1,  
                                                             n_estimators=100,  
                                                             random_seed=0,  
                                                             n_jobs=-1,  
                                                             **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 -1.

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

transformers**Subpackages****dimensionality_reduction****Submodules****lda****Module Contents****Classes Summary**

<i>LinearDiscriminantAnalysis</i>	Reduces the number of features by using Linear Discriminant Analysis.
-----------------------------------	-----------------------------------------------------------------------

Contents

class evalml.pipelines.components.transformers.dimensionality_reduction.lda.**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

continues on next page

Table 380 – continued from previous page

<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

<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>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

continues on next page

Table 384 – 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*)

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

<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*

encoders

Submodules

onehot_encoder

Module Contents

Classes Summary

OneHotEncoder

A transformer that encodes categorical features in a one-hot numeric array.

continues on next page

Table 386 – continued from previous page

<i>OneHotEncoderMeta</i>	A version of the <code>ComponentBaseMeta</code> 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=None, 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

<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.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

<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*)

target_encoder

Module Contents

Classes Summary

*TargetEncoder*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

<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*

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

<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.transformers.encoders.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
<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

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

continues on next page

Table 395 – 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 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

rf_classifier_feature_selector

Module Contents

Classes Summary

RFClassifierSelectFromModel

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

<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*

rf_regressor_feature_selector

Module Contents

Classes Summary

<i>RFRegressorSelectFromModel</i>	Selects top features based on importance weights using a Random Forest regressor.
-----------------------------------	-----------------------------------------------------------------------------------

Contents

class evalml.pipelines.components.transformers.feature_selection.rf_regressor_feature_selector

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`

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

continues on next page

Table 401 – continued from previous page

<code>get_names</code>	Get names of selected features.
<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 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

continues on next page

Table 402 – continued from previous page

<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*

```
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

continues on next page

Table 403 – continued from previous page

<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*

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=cat-  
                                                                    e-  
                                                                    gor-  
                                                                    i-  
                                                                    cal_fill_value=None,  
                                                                    nu-  
                                                                    meric_impute_strategy='me  
                                                                    nu-  
                                                                    meric_fill_value=None,  
                                                                    ran-  
                                                                    dom_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*

per_column_imputer

Module Contents

Classes Summary

PerColumnImputer

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="most_frequent", fill_value=0, random_seed=None, **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*

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_strategy*, *fill_value*, *random_seed*, ***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

continues on next page

Table 411 – continued from previous page

<code>fit_transform</code>	Fits on and transforms the input target 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>save</code>	Saves component at file path
<code>transform</code>	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 416 – 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 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.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

DateTimeFeaturizer

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.
- **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

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`

drop_null_columns

Module Contents

Classes Summary

DropNullColumns

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
<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*

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** (*index*, *random_seed*, *hyperparameter_ranges*, *model_family*, *modifies_features*, *modifies_target*, *name*)

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

hyperparameter_ranges	{}
model_family	ModelFamily.NONE
modifies_features	True
modifies_target	False
name	DFS 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 the DFSTransformer Transformer component.
<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>	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*

lsa

Module Contents

Classes Summary

LSA

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
<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`

polynomial_detrender

Module Contents

Classes Summary

<i>PolynomialDetrender</i>	Removes trends from time series by fitting a polynomial to the data.
----------------------------	----------------------------------------------------------------------

Contents

`class evalml.pipelines.components.transformers.preprocessing.polynomial_detrender.Polynomial`

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.

continues on next page

Table 429 – continued from previous page

<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

<i>TextFeaturizer</i>	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.

Parameters **random_seed** (*int*) – Seed for the random number generator. Defaults to 0.

Attributes

hyper-paramer-ter_ranges	{}
model_family	ModelFamily.NONE
modi-fies_features	True
modi-fies_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*

text_transformer

Module Contents

Classes Summary

<i>TextTransformer</i>	Base class for all transformers working with text features.
------------------------	-------------------------------------------------------------

Attributes Summary

<i>logger</i>

Contents

`evalml.pipelines.components.transformers.preprocessing.text_transformer.logger`

class `evalml.pipelines.components.transformers.preprocessing.text_transformer.TextTransformer`

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

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>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.
<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.

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.
<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.transformers.preprocessing.DelayedFeatureTransformer(*date_index*, *max_delay*, *delay_features*, *delay_target*, *gap*, *random_seed*, ***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.
- **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

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

continues on next page

Table 437 – 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 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', *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

<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 DFSTransformer Transformer component.
<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>	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=*
ran-
dom_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.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

<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 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,  
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.preprocessing.**TextFeaturizer** (*random_seed*=0, ***kwargs*)

Transformer that can automatically featurize text columns using featuretools' nlp_primitives.

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

continues on next page

Table 442 – 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 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

<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 443 – continued from previous page

<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`

samplers

Submodules

base_sampler

Module Contents

Classes Summary

<code>BaseOverSampler</code>	Base Oversampler component. Used as the base class of all imbalance-learn oversampler components.
<code>BaseSampler</code>	Base Sampler component. Used as the base class of all sampler components.

Contents

class evalml.pipelines.components.transformers.samplers.base_sampler.**BaseOverSampler** (*sampler*, *sampling_ratio*, *sampling_ratio_dict*, *k_neighbors_default*, *n_jobs*, *random_seed*, ***kwargs*)

Base Oversampler component. Used as the base class of all imbalance-learn oversampler components.

Parameters

- **sampler** (*obj*) – Sampler object to use.
- **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

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 Oversampler to the data.

continues on next page

Table 445 – continued from previous page

<code>fit_transform</code>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
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

<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>	Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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*)

Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type *pd.DataFrame*, *pd.Series*

oversamplers

Module Contents

Classes Summary

<i>SMOTENCSampler</i>	SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns.
<i>SMOTENSampler</i>	SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets.

continues on next page

Table 447 – continued from previous page

<i>SMOTESampler</i>	SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.
---------------------	-----------------------------------------------------------------------------------------------------------------------------------

Contents

class evalml.pipelines.components.transformers.samplers.oversamplers.**SMOTENC****Sampler** (*sampling_ratio*, *k_neighbors_default*, *n_jobs*=-1, *random_seed*=0, ***kwargs*)

SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns. Input data must be Woodwork type, and this component is only run during training and not during predict.

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.
- **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	SMOTENC 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 Oversampler to the data.

continues on next page

Table 448 – continued from previous page

<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
class evalml.pipelines.components.transformers.samplers.oversamplers.SMOTENSampler (sampling_ratio,  
                                         k_neighbors_default,  
                                         n_jobs=-1,  
                                         random_seed=None,  
                                         **kwargs)
```

SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets. This component is only run during training and not during predict.

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.
- **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	SMOTEN 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self, X, y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

class evalml.pipelines.components.transformers.samplers.oversamplers.**SMOTESampler** (*sampling_rate, k_neighbors, n_jobs=-1, random_seed=0, **kwargs*)

SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.

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.
- **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	SMOTE 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.

- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type *pd.DataFrame*, *pd.Series*

undersampler

Module Contents

Classes Summary

<i>Undersampler</i>	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_dict*, *min_samples*, *min_percentage*, *random_seed*, ***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>	Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.
<i>fit_transform</i>	Fit and transform the data using the undersampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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*)

Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self***fit_transform** (*self, X, y*)

Fit and transform the data using the undersampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.**Return type** *pd.DataFrame, pd.Series*

Package Contents

Classes Summary

<i>SMOTENCSampler</i>	SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns.
<i>SMOTENSampler</i>	SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets.
<i>SMOTESampler</i>	SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.
<i>Undersampler</i>	Initializes an undersampling transformer to downsample the majority classes in the dataset.

Contents

class evalml.pipelines.components.transformers.samplers.**SMOTENCSampler** (*sampling_ratio=0.25*,
k_neighbors_default=5,
n_jobs=-1,
random_seed=0,
***kwargs*)

SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns. Input data must be Woodwork type, and this component is only run during training and not during predict.

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.
- **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	SMOTENC 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self***fit_transform** (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.**Return type** *pd.DataFrame*, *pd.Series*

```
class evalml.pipelines.components.transformers.samplers.SMOTENSampler (sampling_ratio=0.25,  
                                                                    k_neighbors_default=5,  
                                                                    n_jobs=-  
                                                                    1,  
                                                                    ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets. This component is only run during training and not during predict.

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.
- **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	SMOTEN 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
class evalml.pipelines.components.transformers.samplers.SMOTESampler (sampling_ratio=0.25,  
                                                                    k_neighbors_default=5,  
                                                                    n_jobs=-  
                                                                    1, ran-  
                                                                    dom_seed=0,  
                                                                    **kwargs)
```

SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.

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.
- **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	SMOTE 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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 456 – continued from previous page

<i>transform</i>	No transformation needs to be done here.
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</i>)	
Fits the Oversampler to the data.	
Parameters	
<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i>) – Training features • y (<i>pd.Series</i>) – Target features 	
Returns self	
fit_transform (<i>self</i> , <i>X</i> , <i>y</i>)	
Fit and transform the data using the data sampler. Used during training of the pipeline	
Parameters	
<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i>) – Training features • y – Target features 	
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	

save (*self*, *file_path*, *pickle_protocol*=*cloudpickle.DEFAULT_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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

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>	Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.
<i>fit_transform</i>	Fit and transform the data using the undersampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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*)

Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self***fit_transform** (*self, X, y*)

Fit and transform the data using the undersampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.**Return type** *pd.DataFrame, pd.Series*

scalers

Submodules

standard_scaler

Module 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.standard_scaler.**StandardScaler** (*random* ***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

StandardScaler

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

<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

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>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.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

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

continues on next page

Table 467 – continued from previous page

<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

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>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>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>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>SelectColumns</i>	Selects specified columns in input data.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>SMOTENCSampler</i>	SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns.
<i>SMOTENSampler</i>	SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets.
<i>SMOTESampler</i>	SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.
<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 469 – 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.

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
<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 470 – 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.DelayedFeatureTransformer (date_index=None,
                                                                    max_delay=2,
                                                                    de-
                                                                    lay_features=True,
                                                                    de-
                                                                    lay_target=True,
                                                                    gap=1,
                                                                    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.

- **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

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 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.transformers.**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 <i>X</i> and transforms <i>X</i>
<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 <i>X</i> 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,  
                                                             ran-  
                                                             dom_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- parame- ter_ranges	{}
model_family	ModelFamily.NONE
modi- fies_features	True
modi- fies_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.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

<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.

continues on next page

Table 475 – continued from previous page

<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 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.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

<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.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.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 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 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, fea-  

tures_to_encode=None,  

categories=None,  

drop='if_binary',  

han-  

dle_unknown='ignore',  

han-  

dle_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`

[illegible]

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

continues on next page

Table 480 – 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*

```
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*,
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.transformers.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.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

<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*

```
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	{}
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.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

<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.transformers.SMOTENCSampler (sampling_ratio=0.25,  

k_neighbors_default=5,  

n_jobs=- 1, ran-  

dom_seed=0,  

**kwargs)
```

SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns. Input data must be Woodwork type, and this component is only run during training and not during predict.

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.
- **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	SMOTENC 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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 487 – 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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type *pd.DataFrame*, *pd.Series*

```
class evalml.pipelines.components.transformers.SMOTENSampler (sampling_ratio=0.25,
                                                             k_neighbors_default=5,
                                                             n_jobs=- 1, ran-
                                                             dom_seed=0,
                                                             **kwargs)
```

SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets. This component is only run during training and not during predict.

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.
- **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	SMOTEN Oversampler

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 Oversampler to the data.
<code>fit_transform</code>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features

- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
class evalml.pipelines.components.transformers.SMOTESampler (sampling_ratio=0.25,  
                                                         k_neighbors_default=5,  
                                                         n_jobs=- 1, ran-  
                                                         dom_seed=0,  
                                                         **kwargs)
```

SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.

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.
- **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	SMOTE 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self*

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type *pd.DataFrame*, *pd.Series*

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-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

```
class evalml.pipelines.components.transformers.TargetEncoder (cols=None,
                                                             smooth-
                                                             ing=1.0,      han-
                                                             dle_unknown='value',
                                                             han-
                                                             dle_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’, nd ‘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.
<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 491 – 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*)

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 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.TextFeaturizer` (*random_seed=0*,
***kwargs*)

Transformer that can automatically featurize text columns using `featuretools`’ `nlp_primitives`.

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.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 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.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>	Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.
<i>fit_transform</i>	Fit and transform the data using the undersampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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*)

Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self***fit_transform** (*self, X, y*)

Fit and transform the data using the undersampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.**Return type** *pd.DataFrame, pd.Series*

Submodules

component_base

Module Contents

Classes Summary

<i>ComponentBase</i>	Base class for all components.
----------------------	--------------------------------

Attributes Summary

<i>logger</i>

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

continues on next page

Table 498 – 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

clone (*self*)

Constructs a new component with the same parameters 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

`evalml.pipelines.components.component_base.logger`

component_base_meta

Module Contents

Classes Summary

ComponentBaseMeta

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 Component-Base 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>	

Attributes Summary

<i>logger</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 is provided, will return that without modification.

Parameters *component* (*str*, *ComponentBase*) – input to be standardized

Returns ComponentBase

`evalml.pipelines.components.utils.logger`

`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

<code>fit</code>	Fits component to data
<code>get_params</code>	Get parameters for this estimator.
<code>predict</code>	Make predictions using selected features.
<code>predict_proba</code>	Make probability estimates for labels.
<code>score</code>	Return the mean accuracy on the given test data and labels.
<code>set_params</code>	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 $n_{\text{samples_fitted}}$ 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

<code>ARIMAREgressor</code>	Autoregressive Integrated Moving Average Model.
<code>BaselineClassifier</code>	Classifier that predicts using the specified strategy.
<code>BaselineRegressor</code>	Baseline regressor that uses a simple strategy to make predictions.
<code>CatBoostClassifier</code>	CatBoost Classifier, a classifier that uses gradient-boosting on decision trees.

continues on next page

Table 506 – continued from previous page

<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>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>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>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>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>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>SelectColumns</i>	Selects specified columns in input data.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.

continues on next page

Table 506 – continued from previous page

<i>SMOTENCSampler</i>	SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns.
<i>SMOTENSampler</i>	SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets.
<i>SMOTESampler</i>	SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.
<i>StackedEnsembleClassifier</i>	Stacked Ensemble Classifier.
<i>StackedEnsembleRegressor</i>	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>TargetImputer</i>	Imputes missing target data according to a specified imputation strategy.
<i>TextFeaturizer</i>	Transformer that can automatically featurize text columns using featuretools' <code>nlp_primitives</code> .
<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>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
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.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

<code>classes_</code>	Returns class labels. Will return None before fitting.
<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. Since baseline classifiers do not use input features to calculate predictions, returns an array of zeroes.
<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 508 – 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

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
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. Since baseline regressors 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

clone (*self*)

Constructs a new component with the same parameters 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*, 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.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

continues on next page

Table 510 – 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.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
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.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

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

<code>check_for_fit</code>	<code>check_for_fit</code> wraps a method that validates if <code>self._is_fitted</code> is <code>True</code> .
<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*)

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

hyperparameter_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*

```
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 $\text{int}(\text{max_features} * \text{n_features})$ 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.

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

<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.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

continues on next page

Table 516 – continued from previous page

<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.DelayedFeatureTransformer (date_index=None,
                                                             max_delay=2, de-
                                                             lay_features=True,
                                                             delay_target=True,
                                                             gap=1, 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.
- **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

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.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.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.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.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	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
pre-dict_uses_y	<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

continues on next page

Table 523 – 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>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.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 $\text{int}(\text{max_features} * \text{n_features})$ 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
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

continues on next page

Table 524 – 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.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

<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	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<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 component to data
<code>fit_transform</code>	Fits on <i>X</i> and transforms <i>X</i>
<code>get_names</code>	Get names of selected features.
<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

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 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.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

<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 <code>np.nan</code> 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 <code>np.nan</code> 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
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.LightGBMClassifier(boosting_type='gbdt',
                                                    learning_rate=0.1,
                                                    n_estimators=100,
                                                    max_depth=0, num_leaves=31,
                                                    min_child_samples=20, bag-
                                                    gging_fraction=0.9, bag-
                                                    gging_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.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 532 – 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, ran-  
                                                                dom_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.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	<code>{}</code>
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
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 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 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

<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.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	<code>Real(0.25, 1)}</code> :type: {“variance”}
model_family	<code>ModelFamily.NONE</code>
modifies_features	<code>True</code>
modifies_target	<code>False</code>
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*

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

<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 imputers on input data
<code>fit_transform</code>	Fits on X and transforms X

continues on next page

Table 537 – 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 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

<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.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

<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 540 – 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.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.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 <code>np.nan</code> 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 <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.SMOTENCSampler (sampling_ratio=0.25,
                                                k_neighbors_default=5, n_jobs=- 1,
                                                random_seed=0, **kwargs)
```

SMOTENC Oversampler component. Uses SMOTENC to generate synthetic samples. Works on a mix of numerical and categorical columns. Input data must be Woodwork type, and this component is only run during training and not during predict.

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.
- **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	SMOTENC 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 545 – continued from previous page

<i>fit</i>	Fits the Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
class evalml.pipelines.components.SMOTENSampler (sampling_ratio=0.25,
                                              k_neighbors_default=5,    n_jobs=-
                                              1, random_seed=0, **kwargs)
```

SMOTEN Oversampler component. Uses SMOTEN to generate synthetic samples. Works for purely categorical datasets. This component is only run during training and not during predict.

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.
- **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	SMOTEN 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns *self***fit_transform** (*self, X, y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.**Return type** *pd.DataFrame, pd.Series*

```
class evalml.pipelines.components.SMOTESampler (sampling_ratio=0.25,  
                                              k_neighbors_default=5, n_jobs=- 1,  
                                              random_seed=0, **kwargs)
```

SMOTE Oversampler component. Works on numerical datasets only. This component is only run during training and not during predict.

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.

- **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	SMOTE 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 Oversampler to the data.
<i>fit_transform</i>	Fit and transform the data using the data sampler. Used during training of the pipeline
<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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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 Oversampler to the data.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self

fit_transform (*self*, *X*, *y*)

Fit and transform the data using the data sampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns X and y data that was passed in.

Return type pd.DataFrame, pd.Series

```
class evalml.pipelines.components.StackedEnsembleClassifier (input_pipelines=None,
                                                         fi-
                                                         nal_estimator=None,
                                                         cv=None, n_jobs=-
                                                         1, random_seed=0,
                                                         **kwargs)
```

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	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.
--------------	-----------------------------------------------------------------------

continues on next page

Table 548 – continued from previous page

<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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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.StackedEnsembleRegressor (input_pipelines=None,
                                                         fi-
                                                         nal_estimator=None,
                                                         cv=None, n_jobs=-
                                                         1, random_seed=0,
                                                         **kwargs)
```

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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

<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.SVMClassifier(C=1.0, kernel='rbf', gamma='scale',
                                              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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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

<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.

continues on next page

Table 551 – 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*)

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='scale'*, *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** ({*"linear"*, *"poly"*, *"rbf"*, *"sigmoid"*, *"precomputed"*}) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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

<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.**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
<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 <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.**TextFeaturizer** (*random_seed*=0, ***kwargs*)

Transformer that can automatically featurize text columns using featuretools’ nlp_primitives.

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*, *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.
- **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
pre-dict_uses_y	True
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

continues on next page

Table 556 – continued from previous page

<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*, *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*, *y=None*)

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

<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

```
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>	Resample the data using the sampler. Since our sampler doesn’t need to be fit, we do nothing here.
<i>fit_transform</i>	Fit and transform the data using the undersampler. Used during training of the pipeline
<i>load</i>	Loads component at file path

continues on next page

Table 558 – 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>	No transformation needs to be done here.

clone (*self*)

Constructs a new component with the same parameters 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*)

Resample the data using the sampler. Since our sampler doesn't need to be fit, we do nothing here.

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** (*pd.Series*) – Target features

Returns self**fit_transform** (*self*, *X*, *y*)

Fit and transform the data using the undersampler. Used during training of the pipeline

Parameters

- **X** (*pd.DataFrame*) – Training features
- **y** – Target features

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*)

No transformation needs to be done here.

Parameters

- **X** (*pd.DataFrame*) – Training features. Ignored.
- **y** (*pd.Series*) – Target features. Ignored.

Returns *X* and *y* data that was passed in.

Return type `pd.DataFrame`, `pd.Series`

```
class evalml.pipelines.components.XGBoostClassifier(eta=0.1, max_depth=6,
                                                    min_child_weight=1,
                                                    n_estimators=100, random_seed=0, n_jobs=-1,
                                                    **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 -1.

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.XGBoostRegressor(eta=0.1, max_depth=6,
                                                    min_child_weight=1,
                                                    n_estimators=100, random_seed=0, n_jobs=-1,
                                                    **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 -1.

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
<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 560 – 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

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.
<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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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

continues on next page

Table 562 – continued from previous page

<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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**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**property linearized_component_graph** (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

Returns Estimated labels

Return type pd.Series

predict_proba (*self, X*)

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]

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*)

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

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.

binary_classification_pipeline_mixin

Module Contents

Classes Summary

BinaryClassificationPipelineMixin

Contents

class evalml.pipelines.binary_classification_pipeline_mixin.**BinaryClassificationPipelineMixin**

Methods

<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>threshold</i>	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

<i>ClassificationPipeline</i>	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.

continues on next page

Table 566 – continued from previous page

<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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

Returns Estimated labels

Return type `pd.Series`

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters *X* (`pd.DataFrame` or `np.ndarray`) – Data of shape `[n_samples, n_features]`

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*)

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

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

component_graph

Module Contents

Classes Summary

ComponentGraph

Component graph for a pipeline as a directed acyclic graph (DAG).

Attributes Summary

logger

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>from_list</code>	Constructs a linear ComponentGraph from a given list, where each component in the list feeds its X transformed output to the next component
<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_last_component</code>	Retrieves the component that is computed last in the graph, usually the final estimator.
<code>get_parents</code>	Finds all of the inputs for a given component, including the names of all parent nodes of the given component
<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.

continues on next page

Table 569 – continued from previous page

<code>linearized_component_graph</code>	Return a list of (component name, component class) tuples from a pre-initialized component graph defined
<code>predict</code>	Make predictions using selected features.

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 from_list (*cls*, *component_list*, *random_seed=0*)

Constructs a linear ComponentGraph from a given list, where each component in the list feeds its X transformed output to the next component

Parameters `component_list` (*list*) – String names or ComponentBase subclasses in an order that represents a valid linear graph

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_last_component (*self*)

Retrieves the component that is computed last in the graph, usually the final estimator.

Returns ComponentBase object

get_parents (*self*, *component_name*)

Finds all of the inputs for a given component, including the names of all parent nodes of the given component

Parameters `component_name` (*str*) – Name of the child component to look up

Returns List of inputs to use

Return type list[str]

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.

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

classmethod `linearized_component_graph` (*cls*, *components*)

Return a list of (component name, component class) tuples from a pre-initialized component graph defined as either a list or a dictionary. The component names are guaranteed to be unique.

Parameters **components** (*list(ComponentBase) or Dict[str, ComponentBase]*) – Components in the pipeline.

Returns

list((component name, ComponentBase)) - tuples with the unique component name as the first element and the component class as the second element. When the input is a list, the components will be returned in the order they appear in the input.

predict (*self, X*)

Make predictions using selected features.

Parameters **X** (*pd.DataFrame*) – Data of shape [n_samples, n_features]

Returns Predicted values.

Return type *pd.Series*

`evalml.pipelines.component_graph.logger`

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.
<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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

pipeline_base

Module Contents

Classes Summary

PipelineBase

Machine learning pipeline made out of transformers and a 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 a 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 574 – continued from previous page

<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>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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.

`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*)

Transforms the data by applying all pre-processing components.

Parameters **`X`** (`pd.DataFrame`) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

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*)

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

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

pipeline_meta

Module Contents

Classes Summary

<i>PipelineBaseMeta</i>	Metaclass that overrides creating a new pipeline by wrapping methods with validators and setters
<i>TimeSeriesPipelineBaseMeta</i>	Metaclass that overrides creating a new time series 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*)

class evalml.pipelines.pipeline_meta.**TimeSeriesPipelineBaseMeta**

Metaclass that overrides creating a new time series 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

<code>check_for_fit</code>	<code>check_for_fit</code> wraps a method that validates if <code>self.is_fitted</code> is <code>True</code> .
<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*)

regression_pipeline

Module Contents

Classes Summary

<code>RegressionPipeline</code>	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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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*

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

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*)

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

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

time_series_classification_pipelines

Module Contents

Classes Summary

<code>TimeSeriesBinaryClassificationPipeline</code>	Pipeline base class for time series binary classification problems.
<code>TimeSeriesClassificationPipeline</code>	Pipeline base class for time series classification problems.
<code>TimeSeriesMulticlassClassificationPipeline</code>	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>graph</i>	Generate an image representing the pipeline graph

continues on next page

Table 581 – 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>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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.
<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.

`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*)

Transforms the data by applying all pre-processing components.

Parameters ***X*** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

Returns Predicted values.

Return type pd.Series

predict_proba (*self*, *X*, *y=None*)

Make probability estimates for labels.

Parameters *X* (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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.

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

<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>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 <code>inverse_transform</code> methods to estimator predictions in reverse order.
<i>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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 current and additional objectives.

continues on next page

Table 582 – continued from previous page

<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.
can_tune_threshold_with_objective (<i>self</i> , <i>objective</i>)	
Determine whether the threshold of a binary classification pipeline can be tuned.	
Parameters	
<ul style="list-style-type: none"> • pipeline (<i>PipelineBase</i>) – Binary classification pipeline. • objective – Primary AutoMLSearch objective. 	
property classes_ (<i>self</i>)	
Gets the class names for the problem.	
clone (<i>self</i>)	
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 (<i>self</i> , <i>X</i> , <i>y=None</i>)	
Transforms the data by applying all pre-processing components.	
Parameters <i>X</i> (<i>pd.DataFrame</i>) – Input data to the pipeline to transform.	
Returns New transformed features.	
Return type <i>pd.DataFrame</i>	
static create_objectives (<i>objectives</i>)	
property custom_name (<i>self</i>)	
Custom name of the pipeline.	
describe (<i>self</i> , <i>return_dict=False</i>)	
Outputs pipeline details including component parameters	
Parameters return_dict (<i>bool</i>) – If True, return dictionary of information about pipeline. Defaults to False.	
Returns Dictionary of all component parameters if <i>return_dict</i> is True, else None	
Return type <i>dict</i>	
property feature_importance (<i>self</i>)	
Importance associated with each feature. Features dropped by the feature selection are excluded.	
Returns <i>pd.DataFrame</i> including feature names and their corresponding importance	
fit (<i>self</i> , <i>X</i> , <i>y</i>)	
Fit a time series classification pipeline.	
Parameters	
<ul style="list-style-type: none"> • X (<i>pd.DataFrame</i> or <i>np.ndarray</i>) – The input training data of shape [n_samples, n_features] • y (<i>pd.Series</i>, <i>np.ndarray</i>) – The target training targets of length [n_samples] 	
Returns <i>self</i>	
get_component (<i>self</i> , <i>name</i>)	
Returns component by name	

Parameters **name** (*str*) – Name of component

Returns Component to return

Return type Component

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

Returns Predicted values.

Return type *pd.Series*

predict_proba (*self*, *X*, *y=None*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

```
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>graph</i>	Generate an image representing the pipeline graph
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance

continues on next page

Table 583 – continued from previous page

<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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 current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

Returns Predicted values.

Return type pd.Series

predict_proba (*self*, *X*, *y=None*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

time_series_regression_pipeline

Module Contents

Classes Summary

<i>TimeSeriesRegressionPipeline</i>	Pipeline base class for time series regression problems.
-------------------------------------	----------------------------------------------------------

Contents

class evalml.pipelines.time_series_regression_pipeline.**TimeSeriesRegressionPipeline** (*component*

*pa-
ram-
e-
ters=Non
cus-
tom_nam
ran-
dom_see*

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_REGRESSIO
---------------------	------------------------------------

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 regression pipeline.
<i>get_component</i>	Returns component by name
<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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters *X* (*pd.DataFrame*) – Input data to the pipeline to transform.

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 regression 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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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, y=None, objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape `[n_samples, n_features]`
- **y** (*pd.Series, np.ndarray, None*) – The target training targets of length `[n_samples]`
- **objective** (*Object or string*) – The objective to use to make predictions

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*)
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

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

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,

Attributes Summary

<code>logger</code>	
---------------------	--

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*)

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** – The name of the sampler component to add to the pipeline. Only used in classification problems. Defaults to None

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.

continues on next page

Table 588 – continued from previous page

<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 a estimator.
<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.
<i>SimpleImputer</i>	Imputes missing data according to a specified imputation strategy.
<i>StackedEnsembleClassifier</i>	Stacked Ensemble Classifier.
<i>StackedEnsembleRegressor</i>	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

<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 array of 0's with a length of 1 as feature_importance is not defined for ARIMA regressor.
<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 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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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>	Make predictions using selected features.
<i>predict_proba</i>	Make probability estimates for labels. Assumes that the column at index 1 represents the positive label case.
<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.

continues on next page

Table 590 – continued from previous page

<i>threshold</i>	Threshold used to make a prediction. Defaults to None.
------------------	--------------------------------------------------------

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*)
Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

Returns Estimated labels

Return type *pd.Series*

predict_proba (*self*, *X*)

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]

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*)

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

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.

```
class evalml.pipelines.CatBoostClassifier(n_estimators=10, eta=0.03, max_depth=6,
                                          bootstrap_type=None, silent=True, al-
                                          low_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

<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.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
<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.**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>graph</i>	Generate an image representing the pipeline graph
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance

continues on next page

Table 593 – continued from previous page

<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (`pd.DataFrame`) – Input data to the pipeline to transform.

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

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

property `linearized_component_graph` (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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)

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

Returns Estimated labels

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

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

<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>from_list</code>	Constructs a linear ComponentGraph from a given list, where each component in the list feeds its X transformed output to the next component
<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_last_component</code>	Retrieves the component that is computed last in the graph, usually the final estimator.

continues on next page

Table 594 – continued from previous page

<i>get_parents</i>	Finds all of the inputs for a given component, including the names of all parent nodes of the given component
<i>graph</i>	Generate an image representing the component graph
<i>instantiate</i>	Instantiates all uninstantiated components within the graph using the given parameters. An error will be
<i>inverse_transform</i>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<i>linearized_component_graph</i>	Return a list of (component name, component class) tuples from a pre-initialized component graph defined
<i>predict</i>	Make predictions using selected features.

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 **from_list** (*cls, component_list, random_seed=0*)

Constructs a linear ComponentGraph from a given list, where each component in the list feeds its X transformed output to the next component

Parameters **component_list** (*list*) – String names or ComponentBase subclasses in an order that represents a valid linear graph

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_last_component (*self*)

Retrieves the component that is computed last in the graph, usually the final estimator.

Returns ComponentBase object

get_parents (*self, component_name*)

Finds all of the inputs for a given component, including the names of all parent nodes of the given component

Parameters **component_name** (*str*) – Name of the child component to look up

Returns List of inputs to use

Return type *list[str]*

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.

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

classmethod linearized_component_graph (*cls*, *components*)

Return a list of (component name, component class) tuples from a pre-initialized component graph defined as either a list or a dictionary. The component names are guaranteed to be unique.

Parameters *components* (*list(ComponentBase)* or *Dict[str, ComponentBase]*) – Components in the pipeline.

Returns

list((component name, ComponentBase)) - tuples with the unique component name as the first element and the component class as the second element. When the input is a list, the components will be returned in the order they appear in the input.

predict (*self*, *X*)

Make predictions using selected features.

Parameters *X* (*pd.DataFrame*) – Data of shape [n_samples, n_features]

Returns Predicted values.

Return type pd.Series

```
class evalml.pipelines.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
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.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, 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,  
                                                delay_features=True, de-  
                                                lay_target=True, gap=1, 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.
- **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

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

continues on next page

Table 597 – 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 598 – 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 $\text{max_features} = \sqrt{\text{n_features}}$.
 - 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.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 610 – 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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **x** (`pd.DataFrame`) – Input data to the pipeline to transform.

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

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

property `linearized_component_graph` (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

Returns Estimated labels

Return type pd.Series

predict_proba (*self*, *X*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

```
class evalml.pipelines.OneHotEncoder (top_n=10, features_to_encode=None, cat-  
                                         egories=None, drop='if_binary', han-  
                                         dle_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.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.

continues on next page

Table 612 – continued from previous page

<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 a 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

<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 model
<code>get_component</code>	Returns component by name
<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>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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.

`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*)

Transforms the data by applying all pre-processing components.

Parameters **x** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

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*)

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

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

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
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

continues on next page

Table 614 – 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.**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
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.**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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<i>name</i>	Name of the pipeline.

continues on next page

Table 616 – continued from previous page

<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*)

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

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*)

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

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

```
class evalml.pipelines.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.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

<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.StackedEnsembleClassifier (input_pipelines=None, final_estimator=None, cv=None,  
                                              n_jobs=-1, random_seed=0,  
                                              **kwargs)
```

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 StackedEnsembleClassifier and StackedEnsembleRegressor

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.StackedEnsembleRegressor (input_pipelines=None, final_estimator=None, cv=None,
n_jobs=-1, random_seed=0,
**kwargs)
```

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	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 StackedEnsembleClassifier and StackedEnsembleRegressor
<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 621 – 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 StackedEnsembleClassifier and StackedEnsembleRegressor

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 622 – 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*='scale', *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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses 1 / (n_features * X.var()) as value of gamma - if “auto”, 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": ["linear", "poly", "rbf", "sigmoid", "precomputed"], "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='scale'*, *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** (*{ "linear", "poly", "rbf", "sigmoid", "precomputed" }*) – 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 “scale”. - If gamma=’scale’ (default) is passed then it uses $1 / (n_features * X.var())$ as value of gamma - if “auto”, 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”: [“linear”, “poly”, “rbf”, “sigmoid”, “precomputed”], “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 625 – 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>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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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>	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 current and additional objectives.
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **X** (*pd.DataFrame*) – Input data to the pipeline to transform.

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*

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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, y=None, objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame, or np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series, np.ndarray, None*) – The target training targets of length [n_samples]
- **objective** (*Object or string*) – The objective to use to make predictions

Returns Predicted values.

Return type *pd.Series*

predict_proba (*self, X, y=None*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame or np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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.

```
class evalml.pipelines.TimeSeriesClassificationPipeline(component_graph, pa-  
                                                         rameters=None, cus-  
                                                         tom_name=None, ran-  
                                                         dom_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

prob- lem_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>graph</i>	Generate an image representing the pipeline graph
<i>graph_feature_importance</i>	Generate a bar graph of the pipeline’s feature importance

continues on next page

Table 627 – continued from previous page

<code>inverse_transform</code>	Apply component <code>inverse_transform</code> methods to estimator predictions in reverse order.
<code>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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 current and additional objectives.
<code>summary</code>	A short summary of the pipeline structure, describing the list of components used.

`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*)

Transforms the data by applying all pre-processing components.

Parameters **`X`** (`pd.DataFrame`) – Input data to the pipeline to transform.

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

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

property linearized_component_graph (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

Returns Predicted values.

Return type pd.Series

predict_proba (*self*, *X*, *y=None*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

```
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	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.

continues on next page

Table 628 – 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>	Fit a time series classification pipeline.
<i>get_component</i>	Returns component by name
<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>linearized_component_graph</i>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<i>load</i>	Loads pipeline at file path
<i>model_family</i>	Returns model family of this pipeline template
<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 current and additional objectives.
<i>summary</i>	A short summary of the pipeline structure, describing the list of components used.

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*)

Transforms the data by applying all pre-processing components.

Parameters **x** (*pd.DataFrame*) – Input data to the pipeline to transform.

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

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

property `linearized_component_graph` (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

Returns Predicted values.

Return type pd.Series

predict_proba (*self*, *X*, *y=None*)

Make probability estimates for labels.

Parameters **X** (*pd.DataFrame* or *np.ndarray*) – Data of shape [n_samples, n_features]

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*)

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

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

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_REGRESSIO
---------------------	------------------------------------

Methods

<i>can_tune_threshold_with_objective</i>	Determine whether the threshold of a binary classification pipeline can be tuned.
------------------------------------------	-----------------------------------------------------------------------------------

continues on next page

Table 629 – continued from previous page

<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 regression pipeline.
<code>get_component</code>	Returns component by name
<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>linearized_component_graph</code>	A component graph in list form. Note that this is not guaranteed to be in proper component computation order
<code>load</code>	Loads pipeline at file path
<code>model_family</code>	Returns model family of this pipeline template
<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.

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*)

Transforms the data by applying all pre-processing components.

Parameters **x** (`pd.DataFrame`) – Input data to the pipeline to transform.

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 regression 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

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

property `linearized_component_graph` (*self*)

A component graph in list form. Note that this is not guaranteed to be in proper component computation order

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 template

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*, *y=None*, *objective=None*)

Make predictions using selected features.

Parameters

- **X** (*pd.DataFrame*, or *np.ndarray*) – Data of shape [n_samples, n_features]
- **y** (*pd.Series*, *np.ndarray*, *None*) – The target training targets of length [n_samples]
- **objective** (*Object* or *string*) – The objective to use to make predictions

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*)

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

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

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

continues on next page

Table 630 – continued from previous page

<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.XGBoostClassifier (*eta*=0.1, *max_depth*=6, *min_child_weight*=1, *n_estimators*=100, *random_seed*=0, *n_jobs*=-1, ***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 -1.

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=- 1,
                                         **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 -1.

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
<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

<code>BalancedClassificationSampler</code>	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

<code>SamplerBase</code>	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

<code>fit_resample</code>	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 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.

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

<i>BalancedClassificationSampler</i>	Class for balanced classification downsampler.
<i>SamplerBase</i>	Base class for all custom samplers.
<i>TimeSeriesSplit</i>	Rolling Origin Cross Validation for time series problems.
<i>TrainingValidationSplit</i>	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

<i>fit_resample</i>	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

<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.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**

<i>drop_nan_target_rows</i>	Drops rows in X and y when row in the target y has a value of NaN.
<i>load_data</i>	Load features and target from file.
<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.
<i>target_distribution</i>	Get the target distributions.

Contents

`evalml.preprocessing.utils.drop_nan_target_rows(X, y)`

Drops rows in X and y when row in the target y has a value of NaN.

Parameters

- **X** (*pd.DataFrame*, *np.ndarray*) – Data to transform
- **y** (*pd.Series*, *np.ndarray*) – Target data

Returns Transformed X (and y, if passed in) with rows that had a NaN value removed.

Return type *pd.DataFrame*, *pd.DataFrame*

`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>drop_nan_target_rows</code>	Drops rows in X and y when row in the target y has a value of NaN.
<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.drop_nan_target_rows(X, y)`

Drops rows in X and y when row in the target y has a value of NaN.

Parameters

- **x** (`pd.DataFrame`, `np.ndarray`) – Data to transform
- **y** (`pd.Series`, `np.ndarray`) – Target data

Returns Transformed X (and y, if passed in) with rows that had a NaN value removed.

Return type `pd.DataFrame`, `pd.DataFrame`

`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

<code>ProblemTypes</code>	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

<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.

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 <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.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*, *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

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

Tuner

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

<i>add</i>	Not applicable to grid 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>	Returns parameters from <i>_grid_points</i> 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 674 – 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.

Attributes Summary

<code>logger</code>

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.logger`

`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

<i>classproperty</i>	Allows function to be accessed as a class level property.
----------------------	-----------------------------------------------------------

Functions

<i>convert_to_seconds</i>	Converts a string describing a length of time to its length in seconds.
<i>deprecate_arg</i>	Helper to raise warnings when a deprecated arg is used.
<i>drop_rows_with_nans</i>	Drop rows that have any NaNs in all dataframes or series.
<i>get_importable_subclasses</i>	Get importable subclasses of a base class. Used to list all of our
<i>get_random_seed</i>	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.
<i>get_random_state</i>	Generates a numpy.random.RandomState instance using seed.
<i>import_or_raise</i>	Attempts to import the requested library by name.
<i>is_all_numeric</i>	Checks if the given DataFrame contains only numeric values
<i>jupyter_check</i>	Get whether or not the code is being run in a Ipython environment (such as Jupyter Notebook or Jupyter Lab)
<i>pad_with_nans</i>	Pad the beginning num_to_pad rows with nans.
<i>safe_repr</i>	Convert the given value into a string that can safely be used for repr
<i>save_plot</i>	Saves fig to filepath if specified, or to a default location if not.

Attributes Summary

<i>logger</i>
<i>SEED_BOUNDS</i>

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

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 Jupyter environment (such as Jupyter Notebook or Jupyter Lab)

Parameters *None* –

Returns *True* if Jupyter, *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 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.gen_utils.SEED_BOUNDS`

logger

Module Contents

Functions

`get_logger`

`log_subtitle`

`log_title`

`time_elapsed`

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

`numeric_and_boolean_ww`

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

<i>classproperty</i>	Allows function to be accessed as a class level property.
----------------------	-----------------------------------------------------------

Functions

<i>convert_to_seconds</i>	Converts a string describing a length of time to its length in seconds.
<i>deprecate_arg</i>	Helper to raise warnings when a deprecated arg is used.
<i>drop_rows_with_nans</i>	Drop rows that have any NaNs in all dataframes or series.
<i>get_importable_subclasses</i>	Get importable subclasses of a base class. Used to list all of our
<i>get_logger</i>	
<i>get_random_seed</i>	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.
<i>get_random_state</i>	Generates a numpy.random.RandomState instance using seed.
<i>import_or_raise</i>	Attempts to import the requested library by name.
<i>infer_feature_types</i>	Create a Woodwork structure from the given list, pandas, or numpy input, with specified types for columns.
<i>is_all_numeric</i>	Checks if the given DataFrame contains only numeric values
<i>jupyter_check</i>	Get whether or not the code is being run in a Ipython environment (such as Jupyter Notebook or Jupyter Lab)
<i>log_subtitle</i>	
<i>log_title</i>	
<i>pad_with_nans</i>	Pad the beginning num_to_pad rows with nans.
<i>safe_repr</i>	Convert the given value into a string that can safely be used for repr
<i>save_plot</i>	Saves fig to filepath if specified, or to a default location if not.

Attributes Summary

SEED_BOUNDS

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` is of type `plotly.Figure`. Defaults to 'png'.
- **interactive** (*bool*, *optional*) – If True and `fig` is of type `plotly.Figure`, saves the `fig` as interactive
- **static** (*bool*, *optional*) – If True, saves the `fig` as a static image
- **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.
---------------	-----------------------------------------------------

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, engine=None)
```

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`, 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.
- **engine** (*EngineBase or None*) – The engine instance used to evaluate pipelines. If None, a `SequentialEngine` will be used.

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>describe_pipeline</code>	Describe a pipeline

continues on next page

Table 687 – continued from previous page

<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*

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', **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.

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 Release

- Enhancements
- Fixes
- Changes
- Documentation Changes
- Testing Changes

Warning: Breaking Changes

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](#)
- **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 $\geq 1.4.2$ #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 feedstock. [#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 `pymzq` at less than 22.0.0 till we add support [#1756](#)
- Remove `ProphetRegressor` from main as windows tests were flaky [#1764](#)

- **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](#)

v0.17.0 Dec. 29, 2020

- **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 Featuretools 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 `HighVarianceCVDDataCheck` 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 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 an 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 its 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](#)
- Standarizing 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`, [143](#)
`evalml.automl`, [143](#)
`evalml.automl.automl_algorithm`, [143](#)
`evalml.automl.automl_algorithm.automl_algorithm`,
[144](#)
`evalml.automl.automl_algorithm.iterative_algorithm`,
[145](#)
`evalml.automl.automl_search`, [165](#)
`evalml.automl.callbacks`, [171](#)
`evalml.automl.engine`, [150](#)
`evalml.automl.engine.cf_engine`, [150](#)
`evalml.automl.engine.dask_engine`, [153](#)
`evalml.automl.engine.engine_base`, [155](#)
`evalml.automl.engine.sequential_engine`,
[158](#)
`evalml.automl.pipeline_search_plots`, [172](#)
`evalml.automl.utils`, [173](#)
`evalml.data_checks`, [183](#)
`evalml.data_checks.class_imbalance_data_check`,
[183](#)
`evalml.data_checks.data_check`, [185](#)
`evalml.data_checks.data_check_action`,
[186](#)
`evalml.data_checks.data_check_action_code`,
[187](#)
`evalml.data_checks.data_check_message`,
[187](#)
`evalml.data_checks.data_check_message_code`,
[189](#)
`evalml.data_checks.data_check_message_type`,
[191](#)
`evalml.data_checks.data_checks`, [191](#)
`evalml.data_checks.datetime_nan_data_check`,
[192](#)
`evalml.data_checks.default_data_checks`,
[193](#)
`evalml.data_checks.highly_null_data_check`,
[195](#)
`evalml.data_checks.id_columns_data_check`,
[196](#)
`evalml.data_checks.invalid_targets_data_check`,
[197](#)
`evalml.data_checks.multicollinearity_data_check`,
[199](#)
`evalml.data_checks.natural_language_nan_data_check`,
[199](#)
`evalml.data_checks.no_variance_data_check`,
[201](#)
`evalml.data_checks.outliers_data_check`,
[202](#)
`evalml.data_checks.sparsity_data_check`,
[203](#)
`evalml.data_checks.target_leakage_data_check`,
[204](#)
`evalml.data_checks.uniqueness_data_check`,
[206](#)
`evalml.data_checks.utils`, [207](#)
`evalml.demos`, [226](#)
`evalml.demos.breast_cancer`, [226](#)
`evalml.demos.churn`, [226](#)
`evalml.demos.diabetes`, [227](#)
`evalml.demos.fraud`, [227](#)
`evalml.demos.wine`, [228](#)
`evalml.exceptions`, [229](#)
`evalml.exceptions.exceptions`, [229](#)
`evalml.model_family`, [231](#)
`evalml.model_family.model_family`, [231](#)
`evalml.model_family.utils`, [232](#)
`evalml.model_understanding`, [234](#)
`evalml.model_understanding.force_plots`,
[241](#)
`evalml.model_understanding.graphs`, [242](#)
`evalml.model_understanding.permutation_importance`,
[252](#)
`evalml.model_understanding.prediction_explanations`,
[234](#)
`evalml.model_understanding.prediction_explanations`,
[234](#)
`evalml.objectives`, [265](#)
`evalml.objectives.binary_classification_objective`,
[265](#)
`evalml.objectives.cost_benefit_matrix`,
[268](#)

<code>evalml.objectives.fraud_cost</code> , 270	<code>evalml.pipelines.components.estimators.classifiers</code>
<code>evalml.objectives.lead_scoring</code> , 273	470
<code>evalml.objectives.multiclass_classification_objective</code> , 276	<code>evalml.pipelines.components.estimators.classifiers</code>
<code>evalml.objectives.objective_base</code> , 278	472
<code>evalml.objectives.regression_objective</code> , 280	<code>evalml.pipelines.components.estimators.classifiers</code>
<code>evalml.objectives.sensitivity_low_alert</code> , 283	475
<code>evalml.objectives.standard_metrics</code> , 285	<code>evalml.pipelines.components.estimators.estimator</code> , 577
<code>evalml.objectives.time_series_regression_objective</code> , 345	<code>evalml.pipelines.components.estimators.regressors</code> , 509
<code>evalml.objectives.utils</code> , 347	<code>evalml.pipelines.components.estimators.regressors</code> , 509
<code>evalml.pipelines</code> , 427	<code>evalml.pipelines.components.estimators.regressors</code> , 513
<code>evalml.pipelines.binary_classification_pipeline</code> , 986	<code>evalml.pipelines.components.estimators.regressors</code> , 515
<code>evalml.pipelines.binary_classification_pipeline_base</code> , 991	<code>evalml.pipelines.components.estimators.regressors</code> , 518
<code>evalml.pipelines.classification_pipeline</code> , 992	<code>evalml.pipelines.components.estimators.regressors</code> , 522
<code>evalml.pipelines.component_graph</code> , 996	<code>evalml.pipelines.components.estimators.regressors</code> , 524
<code>evalml.pipelines.components</code> , 427	<code>evalml.pipelines.components.estimators.regressors</code> , 528
<code>evalml.pipelines.components.component_base</code> , 848	<code>evalml.pipelines.components.estimators.regressors</code> , 531
<code>evalml.pipelines.components.component_base</code> , 850	<code>evalml.pipelines.components.estimators.regressors</code> , 534
<code>evalml.pipelines.components.ensemble</code> , 427	<code>evalml.pipelines.components.estimators.regressors</code> , 536
<code>evalml.pipelines.components.ensemble.stacked_ensemble_base</code> , 427	<code>evalml.pipelines.components.estimators.regressors</code> , 539
<code>evalml.pipelines.components.ensemble.stacked_ensemble_classifier</code> , 430	<code>evalml.pipelines.components.estimators.regressors</code> , 542
<code>evalml.pipelines.components.ensemble.stacked_ensemble_regressor</code> , 433	<code>evalml.pipelines.components.transformers</code> , 643
<code>evalml.pipelines.components.estimators</code> , 445	<code>evalml.pipelines.components.transformers.column_selector</code> , 774
<code>evalml.pipelines.components.estimators.classifier</code> , 445	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 643
<code>evalml.pipelines.components.estimators.classifier_base</code> , 445	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 643
<code>evalml.pipelines.components.estimators.classifier.boosted_classifier</code> , 448	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 645
<code>evalml.pipelines.components.estimators.classifier.linear_classifier</code> , 451	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 652
<code>evalml.pipelines.components.estimators.classifier.naive_bayes_classifier</code> , 454	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 652
<code>evalml.pipelines.components.estimators.classifier.neural_network_classifier</code> , 457	<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> , 657
<code>evalml.pipelines.components.estimators.classifier.random_forest_classifier</code> , 460	<code>evalml.pipelines.components.transformers.feature_selector</code> , 666
<code>evalml.pipelines.components.estimators.classifier.support_vector_machine_classifier</code> , 463	<code>evalml.pipelines.components.transformers.feature_selector</code> , 666
<code>evalml.pipelines.components.estimators.classifier.xgboost_classifier</code> , 467	

A

`abs_error()` (in module `evalml.model_understanding.prediction_explanations.explainers`), 235
`AccuracyBinary` (class in `evalml.objectives`), 350
`AccuracyBinary` (class in `evalml.objectives.standard_metrics`), 286
`AccuracyMulticlass` (class in `evalml.objectives`), 352
`AccuracyMulticlass` (class in `evalml.objectives.standard_metrics`), 288
`add()` (`evalml.tuners.grid_search_tuner.GridSearchTuner` method), 1178
`add()` (`evalml.tuners.GridSearchTuner` method), 1183
`add()` (`evalml.tuners.random_search_tuner.RandomSearchTuner` method), 1179
`add()` (`evalml.tuners.RandomSearchTuner` method), 1184
`add()` (`evalml.tuners.skopt_tuner.SKOptTuner` method), 1180
`add()` (`evalml.tuners.SKOptTuner` method), 1185
`add()` (`evalml.tuners.Tuner` method), 1186
`add()` (`evalml.tuners.tuner.Tuner` method), 1181
`add_result()` (`evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm` method), 145
`add_result()` (`evalml.automl.automl_algorithm.AutoMLAlgorithm` method), 148
`add_result()` (`evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm` method), 147
`add_result()` (`evalml.automl.automl_algorithm.IterativeAlgorithm` method), 150
`add_to_rankings()` (`evalml.automl.automl_search.AutoMLSearch` method), 169
`add_to_rankings()` (`evalml.automl.AutoMLSearch` method), 179
`add_to_rankings()` (`evalml.AutoMLSearch` method), 1201
`all_components()` (in module `evalml.pipelines.components.utils`), 852
`all_problem_types()` (`evalml.problem_types.problem_types.ProblemTypes` method), 1172
`all_problem_types()` (`evalml.problem_types.ProblemTypes` method), 1176
`allowed_model_families()` (in module `evalml.pipelines.components.utils`), 852
`ARIMAREgressor` (class in `evalml.pipelines`), 1035
`ARIMAREgressor` (class in `evalml.pipelines.components`), 857
`ARIMAREgressor` (class in `evalml.pipelines.components.estimators`), 580
`ARIMAREgressor` (class in `evalml.pipelines.components.estimators.regressors`), 545
`ARIMAREgressor` (class in `evalml.pipelines.components.estimators.regressors.arima_regressors`), 510
`AUC` (class in `evalml.objectives`), 354
`AUC` (class in `evalml.objectives.standard_metrics`), 290
`AUCMacro` (class in `evalml.objectives`), 356
`AUCMacro` (class in `evalml.objectives.standard_metrics`), 292
`AUCMicro` (class in `evalml.objectives`), 357
`AUCMicro` (class in `evalml.objectives.standard_metrics`), 294
`AUCWeighted` (class in `evalml.objectives`), 359
`AUCWeighted` (class in `evalml.objectives.standard_metrics`), 295
`AutoMLAlgorithm` (class in `evalml.automl.automl_algorithm`), 148
`AutoMLAlgorithm` (class in `evalml.automl.automl_algorithm.automl_algorithm`), 144
`AutoMLAlgorithmException`, 145, 149
`AutoMLConfig` (in module `evalml.automl.utils`), 173
`AutoMLSearch` (class in `evalml`), 1198
`AutoMLSearch` (class in `evalml.automl`), 176
`AutoMLSearch` (class in `evalml.automl.automl_search`), 166
`AutoMLSearchException`, 229, 230

B

- property), 179
- BalancedAccuracyBinary (class in best_pipeline() (evalml.AutoMLSearch property), 1201
- BalancedAccuracyBinary (class in binary_objective_vs_threshold() (in module evalml.model_understanding), 254
- BalancedAccuracyMulticlass (class in binary_objective_vs_threshold() (in module evalml.model_understanding.graphs), 243
- BalancedAccuracyMulticlass (class in BinaryClassificationObjective (class in evalml.objectives), 364
- BalancedClassificationSampler (class in BinaryClassificationObjective (class in evalml.objectives.binary_classification_objective), 265
- BalancedClassificationSampler (class in BinaryClassificationPipeline (class in evalml.pipelines), 1037
- BaselineClassifier (class in BinaryClassificationPipeline (class in evalml.pipelines.binary_classification_pipeline), 986
- BaselineClassifier (class in BinaryClassificationPipelineMixin (class in evalml.pipelines.binary_classification_pipeline_mixin), 991
- BaselineClassifier (class in evalml.pipelines.components.estimators.classifiers), 479
- BaselineClassifier (class in calculate_percent_difference() (evalml.objectives.AccuracyBinary class method), 351
- BaselineRegressor (class in calculate_percent_difference() (evalml.objectives.AccuracyMulticlass class method), 353
- BaselineRegressor (class in calculate_percent_difference() (evalml.objectives.AUC class method), 354
- BaselineRegressor (class in calculate_percent_difference() (evalml.objectives.AUCMacro class method), 356
- BaselineRegressor (class in calculate_percent_difference() (evalml.objectives.AUCMicro class method), 358
- BaseMeta (class in evalml.utils.base_meta), 1187
- BaseOverSampler (class in calculate_percent_difference() (evalml.objectives.AUCWeighted class method), 360
- BaseSampler (class in calculate_percent_difference() (evalml.objectives.BalancedAccuracyBinary class method), 361
- batch_number() (evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm property), 145
- batch_number() (evalml.automl.automl_algorithm.AutoMLAlgorithm property), 148
- batch_number() (evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm property), 147
- batch_number() (evalml.automl.automl_algorithm.IterativeAlgorithm property), 150
- best_pipeline() (evalml.automl.automl_search.AutoMLSearch property), 169
- best_pipeline() (evalml.automl.AutoMLSearch

<code>calculate_percent_difference()</code> (<i>evalml.objectives.CostBenefitMatrix</i> class method), 368	<code>calculate_percent_difference()</code> (<i>evalml.objectives.MSE</i> class method), 399
<code>calculate_percent_difference()</code> (<i>evalml.objectives.ExpVariance</i> class method), 370	<code>calculate_percent_difference()</code> (<i>evalml.objectives.multiclass_classification_objective.MulticlassC</i> class method), 277
<code>calculate_percent_difference()</code> (<i>evalml.objectives.F1</i> class method), 372	<code>calculate_percent_difference()</code> (<i>evalml.objectives.MulticlassClassificationObjective</i> class method), 401
<code>calculate_percent_difference()</code> (<i>evalml.objectives.F1Macro</i> class method), 374	<code>calculate_percent_difference()</code> (<i>evalml.objectives.objective_base.ObjectiveBase</i> class method), 279
<code>calculate_percent_difference()</code> (<i>evalml.objectives.F1Micro</i> class method), 375	<code>calculate_percent_difference()</code> (<i>evalml.objectives.ObjectiveBase</i> class method), 403
<code>calculate_percent_difference()</code> (<i>evalml.objectives.F1Weighted</i> class method), 377	<code>calculate_percent_difference()</code> (<i>evalml.objectives.Precision</i> class method), 405
<code>calculate_percent_difference()</code> (<i>evalml.objectives.fraud_cost.FraudCost</i> class method), 271	<code>calculate_percent_difference()</code> (<i>evalml.objectives.PrecisionMacro</i> class method), 407
<code>calculate_percent_difference()</code> (<i>evalml.objectives.FraudCost</i> class method), 379	<code>calculate_percent_difference()</code> (<i>evalml.objectives.PrecisionMicro</i> class method), 409
<code>calculate_percent_difference()</code> (<i>evalml.objectives.lead_scoring.LeadScoring</i> class method), 274	<code>calculate_percent_difference()</code> (<i>evalml.objectives.PrecisionWeighted</i> class method), 410
<code>calculate_percent_difference()</code> (<i>evalml.objectives.LeadScoring</i> class method), 382	<code>calculate_percent_difference()</code> (<i>evalml.objectives.R2</i> class method), 412
<code>calculate_percent_difference()</code> (<i>evalml.objectives.LogLossBinary</i> class method), 384	<code>calculate_percent_difference()</code> (<i>evalml.objectives.Recall</i> class method), 413
<code>calculate_percent_difference()</code> (<i>evalml.objectives.LogLossMulticlass</i> class method), 386	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RecallMacro</i> class method), 415
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MAE</i> class method), 388	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RecallMicro</i> class method), 417
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MAPE</i> class method), 389	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RecallWeighted</i> class method), 419
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MaxError</i> class method), 391	<code>calculate_percent_difference()</code> (<i>evalml.objectives.regression_objective.RegressionObjective</i> class method), 281
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MCCBinary</i> class method), 392	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RegressionObjective</i> class method), 420
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MCCMulticlass</i> class method), 395	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RootMeanSquaredError</i> class method), 422
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MeanSquaredLogError</i> class method), 396	<code>calculate_percent_difference()</code> (<i>evalml.objectives.RootMeanSquaredLogError</i> class method), 424
<code>calculate_percent_difference()</code> (<i>evalml.objectives.MedianAE</i> class method), 398	<code>calculate_percent_difference()</code>

<code>(evalml.objectives.sensitivity_low_alert.SensitivityLowAlert(evalml.objectives.standard_metrics.MAPE</code>	<code>(evalml.objectives.sensitivity_low_alert.SensitivityLowAlert(evalml.objectives.standard_metrics.MAPE</code>
<code>class method), 284</code>	<code>class method), 315</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.SensitivityLowAlert class</code>	<code>(evalml.objectives.standard_metrics.MaxError</code>
<code>method), 425</code>	<code>class method), 317</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AccuracyBinary</code>	<code>(evalml.objectives.standard_metrics.MCCBinary</code>
<code>class method), 287</code>	<code>class method), 318</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AccuracyMulticlass</code>	<code>(evalml.objectives.standard_metrics.MCCMulticlass</code>
<code>class method), 289</code>	<code>class method), 320</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AUC</code>	<code>(evalml.objectives.standard_metrics.MeanSquaredLogError</code>
<code>class method), 291</code>	<code>class method), 322</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AUCMacro</code>	<code>(evalml.objectives.standard_metrics.MedianAE</code>
<code>class method), 293</code>	<code>class method), 324</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AUCMicro</code>	<code>(evalml.objectives.standard_metrics.MSE</code>
<code>class method), 294</code>	<code>class method), 325</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.AUCWeighted</code>	<code>(evalml.objectives.standard_metrics.Precision</code>
<code>class method), 296</code>	<code>class method), 327</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.BalancedAccuracyBinary</code>	<code>(evalml.objectives.standard_metrics.PrecisionMacro</code>
<code>class method), 298</code>	<code>class method), 329</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.BalancedAccuracyMulticlass</code>	<code>(evalml.objectives.standard_metrics.PrecisionMicro</code>
<code>class method), 300</code>	<code>class method), 331</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.ExpVariance</code>	<code>(evalml.objectives.standard_metrics.PrecisionWeighted</code>
<code>class method), 301</code>	<code>class method), 332</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.F1 class</code>	<code>(evalml.objectives.standard_metrics.R2 class</code>
<code>method), 303</code>	<code>method), 334</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.F1Macro</code>	<code>(evalml.objectives.standard_metrics.Recall</code>
<code>class method), 305</code>	<code>class method), 335</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.F1Micro</code>	<code>(evalml.objectives.standard_metrics.RecallMacro</code>
<code>class method), 307</code>	<code>class method), 337</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.F1Weighted</code>	<code>(evalml.objectives.standard_metrics.RecallMicro</code>
<code>class method), 308</code>	<code>class method), 339</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.LogLossBinary</code>	<code>(evalml.objectives.standard_metrics.RecallWeighted</code>
<code>class method), 310</code>	<code>class method), 341</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.LogLossMulticlass</code>	<code>(evalml.objectives.standard_metrics.RootMeanSquaredError</code>
<code>class method), 312</code>	<code>class method), 342</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>
<code>(evalml.objectives.standard_metrics.MAE</code>	<code>(evalml.objectives.standard_metrics.RootMeanSquaredLogError</code>
<code>class method), 314</code>	<code>class method), 344</code>
<code>calculate_percent_difference()</code>	<code>calculate_percent_difference()</code>

`(evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective)`
`class method`), 346 `(evalml.objectives.Recall property)`, 414
`calculate_permutation_importance()` `(in module evalml.model_understanding)`, 255 `(evalml.objectives.sensitivity_low_alert.SensitivityLowAlert property)`, 284
`calculate_permutation_importance()` `(in module evalml.model_understanding.permutation_importance)`, 252 `(evalml.objectives.SensitivityLowAlert property)`, 426
`calculate_permutation_importance_one_column()` `(in module evalml.model_understanding)`, 255 `can_optimize_threshold()`
`calculate_permutation_importance_one_column()` `(in module evalml.model_understanding.permutation_importance)`, 252 `(evalml.objectives.standard_metrics.AccuracyBinary property)`, 287
`can_optimize_threshold()` `(evalml.objectives.AccuracyBinary property)`, 351 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.AUC property)`, 355 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.BalancedAccuracyBinary property)`, 362 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.binary_classification_objective.BinaryClassificationObjective property)`, 266 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.BinaryClassificationObjective property)`, 365 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.cost_benefit_matrix.CostBenefitMatrix property)`, 269 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.CostBenefitMatrix property)`, 368 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.F1 property)`, 372 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.fraud_cost.FraudCost property)`, 272 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.FraudCost property)`, 379 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.lead_scoring.LeadScoring property)`, 274 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.LeadScoring property)`, 382 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.LogLossBinary property)`, 384 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.MCCBinary property)`, 393 `can_optimize_threshold()`
`can_optimize_threshold()` `(evalml.objectives.Precision property)`, 405 `can_optimize_threshold()`

(*evalml.pipelines.PipelineBase* method), 1103
 can_tune_threshold_with_objective() (class in *evalml.pipelines*), 1044
 (*evalml.pipelines.regression_pipeline.RegressionPipeline* method), 1011
 can_tune_threshold_with_objective() (class in *evalml.pipelines.components*), 867
 (*evalml.pipelines.RegressionPipeline* method), *evalml.pipelines.components.estimators*, 590
 can_tune_threshold_with_objective() (class in *evalml.pipelines.components.estimators.regressors*), 550
 (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline* method), 1016
 can_tune_threshold_with_objective() (class in *evalml.pipelines.components.estimators.regressors.catboost_regressor*), 516
 (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1021
 can_tune_threshold_with_objective() (*evalml.pipelines.components.OneHotEncoder* method), 624
 (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1025
 can_tune_threshold_with_objective() (*evalml.pipelines.components.transformers.encoders.one_hot_encoder* method), 654
 (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* method), 661
 can_tune_threshold_with_objective() (*evalml.pipelines.components.transformers.encoders.OneHotEncoder* method), 808
 (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* method), 1137
 can_tune_threshold_with_objective() (*evalml.pipelines.OneHotEncoder* method), 1098
 (*evalml.pipelines.TimeSeriesClassificationPipeline* method), 1142
 can_tune_threshold_with_objective() (class in *evalml.automl.engine.cf_engine*), 151
 (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* method), 1146
 CFEngine (class in *evalml.automl.engine*), 160
 can_tune_threshold_with_objective() (class in *evalml.automl.engine.cf_engine*), 152
 (*evalml.pipelines.TimeSeriesRegressionPipeline* method), 1150
 check_all_pipeline_names_unique() (in module *evalml.automl.utils*), 173
 cancel() (*evalml.automl.engine.cf_engine.CFComputation* method), 151
 check_for_fit() (*evalml.pipelines.components.component_base_meta* class method), 850
 cancel() (*evalml.automl.engine.dask_engine.DaskComputation* method), 153
 check_for_fit() (*evalml.pipelines.components.ComponentBaseMeta* class method), 872
 cancel() (*evalml.automl.engine.engine_base.EngineComputation* method), 156
 check_for_fit() (*evalml.pipelines.components.transformers.encoders* class method), 656
 cancel() (*evalml.automl.engine.EngineComputation* method), 163
 check_for_fit() (*evalml.pipelines.components.transformers.imputers* class method), 694
 cancel() (*evalml.automl.engine.sequential_engine.SequentialComputation* method), 159
 check_for_fit() (*evalml.pipelines.pipeline_meta.PipelineBaseMeta* class method), 1009
 CatBoostClassifier (class in *evalml.pipelines*), 1042
 check_for_fit() (*evalml.pipelines.pipeline_meta.TimeSeriesPipelineMeta* class method), 1010
 CatBoostClassifier (class in *evalml.pipelines.components*), 865
 classes_() (*evalml.pipelines.binary_classification_pipeline.BinaryClassifier* property), 988
 CatBoostClassifier (class in *evalml.pipelines.components.estimators*), 588
 classes_() (*evalml.pipelines.BinaryClassificationPipeline* property), 1039
 CatBoostClassifier (class in *evalml.pipelines.components.estimators.classifiers*), 481
 classes_() (*evalml.pipelines.classification_pipeline.ClassificationPipeline* property), 993
 CatBoostClassifier (class in *evalml.pipelines.components.estimators.classifiers.catboost_classifier*), 448
 classes_() (*evalml.pipelines.ClassificationPipeline* property), 1048
 (*evalml.pipelines.components.BaselineClassifier* property), 861


```

clone() (evalml.pipelines.components.estimators.classifiers.catboost.CatBoostClassifier.
method), 449
clone() (evalml.pipelines.components.estimators.classifiers.catboost.CatBoostClassifier.
method), 482
clone() (evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier.
method), 452
clone() (evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier.
method), 485
clone() (evalml.pipelines.components.estimators.classifiers.elasticnet.ElasticNetClassifier.
method), 455
clone() (evalml.pipelines.components.estimators.classifiers.elasticnet.ElasticNetClassifier.
method), 488
clone() (evalml.pipelines.components.estimators.classifiers.ensemble.ExtraTreesClassifier.
method), 458
clone() (evalml.pipelines.components.estimators.classifiers.ensemble.ExtraTreesClassifier.
method), 491
clone() (evalml.pipelines.components.estimators.classifiers.kneighbors.KNeighborsClassifier.
method), 462
clone() (evalml.pipelines.components.estimators.classifiers.kneighbors.KNeighborsClassifier.
method), 494
clone() (evalml.pipelines.components.estimators.classifiers.lightgbm.LightGBMClassifier.
method), 465
clone() (evalml.pipelines.components.estimators.classifiers.lightgbm.LightGBMClassifier.
method), 497
clone() (evalml.pipelines.components.estimators.classifiers.logistic.LogisticRegressionClassifier.
method), 468
clone() (evalml.pipelines.components.estimators.classifiers.logistic.LogisticRegressionClassifier.
method), 500
clone() (evalml.pipelines.components.estimators.classifiers.random_forest.RandomForestClassifier.
method), 502
clone() (evalml.pipelines.components.estimators.classifiers.rf_classifier.RandomForestClassifier.
method), 471
clone() (evalml.pipelines.components.estimators.classifiers.svm.SVCClassifier.
method), 474
clone() (evalml.pipelines.components.estimators.classifiers.svm.SVCClassifier.
method), 505
clone() (evalml.pipelines.components.estimators.classifiers.xgboost.XGBoostClassifier.
method), 477
clone() (evalml.pipelines.components.estimators.classifiers.xgboost.XGBoostClassifier.
method), 508
clone() (evalml.pipelines.components.estimators.DecisionTreeClassifier.
method), 594
clone() (evalml.pipelines.components.estimators.DecisionTreeRegressor.
method), 597
clone() (evalml.pipelines.components.estimators.ElasticNetClassifier.
method), 600
clone() (evalml.pipelines.components.estimators.ElasticNetRegressor.
method), 603
clone() (evalml.pipelines.components.estimators.Estimator.
method), 605
clone() (evalml.pipelines.components.estimators.estimator.Estimator.
method), 578
clone() (evalml.pipelines.components.estimators.ExtraTreesClassifier.
method), 608

```

1259

`clone()` (`evalml.pipelines.components.transformers.DropColumns`) (`evalml.pipelines.components.transformers.PolynomialDetrender`
`method`), 794 `method`), 815
`clone()` (`evalml.pipelines.components.transformers.DropNullColumns`) (`evalml.pipelines.components.transformers.preprocessing.datetime`
`method`), 796 `method`), 706
`clone()` (`evalml.pipelines.components.transformers.encoders.one_hot`) (`evalml.pipelines.components.transformers.preprocessing.DateT`
`method`), 654 `method`), 727
`clone()` (`evalml.pipelines.components.transformers.encoders.OneHotEncoder`) (`evalml.pipelines.components.transformers.preprocessing.delaye`
`method`), 661 `method`), 709
`clone()` (`evalml.pipelines.components.transformers.encoders.target_encoder`) (`evalml.pipelines.components.transformers.preprocessing.Delaye`
`method`), 658 `method`), 730
`clone()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder`) (`evalml.pipelines.components.transformers.preprocessing.DFST`
`method`), 664 `method`), 732
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector`) (`evalml.pipelines.components.transformers.preprocessing.drop_`
`method`), 667 `method`), 711
`clone()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`) (`evalml.pipelines.components.transformers.preprocessing.DropN`
`method`), 676 `method`), 734
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector`) (`evalml.pipelines.components.transformers.preprocessing.DropN`
`method`), 670 `method`), 714
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector`) (`evalml.pipelines.components.transformers.preprocessing.DropN`
`method`), 673 `method`), 736
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector`) (`evalml.pipelines.components.transformers.preprocessing.DropN`
`method`), 679 `method`), 716
`clone()` (`evalml.pipelines.components.transformers.feature_selection.feature_selector`) (`evalml.pipelines.components.transformers.preprocessing.DropN`
`method`), 682 `method`), 719
`clone()` (`evalml.pipelines.components.transformers.FeatureSelector`) (`evalml.pipelines.components.transformers.preprocessing.Polyno`
`method`), 799 `method`), 738
`clone()` (`evalml.pipelines.components.transformers.Imputer`) (`evalml.pipelines.components.transformers.preprocessing.text_fe`
`method`), 801 `method`), 721
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.preprocessing.text_tr`
`method`), 696 `method`), 724
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.preprocessing.TextFe`
`method`), 685 `method`), 741
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.preprocessing.TextTr`
`method`), 687 `method`), 743
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.preprocessing.TextTr`
`method`), 698 `method`), 818
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.RFClassifierSelectFr`
`method`), 690 `method`), 821
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.RFRegressorSelectFr`
`method`), 701 `method`), 746
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.samplers.base_samp`
`method`), 693 `method`), 748
`clone()` (`evalml.pipelines.components.transformers.imputer.imputer`) (`evalml.pipelines.components.transformers.samplers.base_samp`
`method`), 703 `method`), 751
`clone()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis`) (`evalml.pipelines.components.transformers.samplers.oversample`
`method`), 803 `method`), 753
`clone()` (`evalml.pipelines.components.transformers.LSA`) (`evalml.pipelines.components.transformers.samplers.oversample`
`method`), 805 `method`), 755
`clone()` (`evalml.pipelines.components.transformers.OneHotEncoder`) (`evalml.pipelines.components.transformers.samplers.SMOTENC`
`method`), 808 `method`), 761
`clone()` (`evalml.pipelines.components.transformers.PCA`) (`evalml.pipelines.components.transformers.samplers.SMOTENS`
`method`), 811 `method`), 763
`clone()` (`evalml.pipelines.components.transformers.PerColumnImputer`) (`evalml.pipelines.components.transformers.samplers.SMOTESan`
`method`), 813 `method`), 766

[clone\(\) \(evalml.pipelines.components.transformers.samplers.UnderSampler method\), 768](#)
[clone\(\) \(evalml.pipelines.components.transformers.samplers.undersampler method\), 758](#)
[clone\(\) \(evalml.pipelines.components.transformers.scalers.standard_scaler.StandardScaler method\), 770](#)
[clone\(\) \(evalml.pipelines.components.transformers.scalers.StandardScaler method\), 773](#)
[clone\(\) \(evalml.pipelines.components.transformers.SelectColumns method\), 823](#)
[clone\(\) \(evalml.pipelines.components.transformers.SimpleImputer method\), 825](#)
[clone\(\) \(evalml.pipelines.components.transformers.SMOTENC sampler method\), 828](#)
[clone\(\) \(evalml.pipelines.components.transformers.SMOTENS sampler method\), 830](#)
[clone\(\) \(evalml.pipelines.components.transformers.SMOTESampler method\), 832](#)
[clone\(\) \(evalml.pipelines.components.transformers.StandardScaler method\), 834](#)
[clone\(\) \(evalml.pipelines.components.transformers.TargetEncoder method\), 837](#)
[clone\(\) \(evalml.pipelines.components.transformers.TargetImputer method\), 839](#)
[clone\(\) \(evalml.pipelines.components.transformers.TextFeaturizer method\), 841](#)
[clone\(\) \(evalml.pipelines.components.transformers.Transformer method\), 843](#)
[clone\(\) \(evalml.pipelines.components.transformers.transformer.TargetTransformer method\), 782](#)
[clone\(\) \(evalml.pipelines.components.transformers.transformer.TargetTransformer method\), 784](#)
[clone\(\) \(evalml.pipelines.components.transformers.UnderSampler method\), 846](#)
[clone\(\) \(evalml.pipelines.components.Undersampler method\), 980](#)
[clone\(\) \(evalml.pipelines.components.XGBoostClassifier method\), 982](#)
[clone\(\) \(evalml.pipelines.components.XGBoostRegressor method\), 985](#)
[clone\(\) \(evalml.pipelines.DecisionTreeClassifier method\), 1055](#)
[clone\(\) \(evalml.pipelines.DecisionTreeRegressor method\), 1058](#)
[clone\(\) \(evalml.pipelines.DelayedFeatureTransformer method\), 1061](#)
[clone\(\) \(evalml.pipelines.DFSTransformer method\), 1063](#)
[clone\(\) \(evalml.pipelines.ElasticNetClassifier method\), 1065](#)
[clone\(\) \(evalml.pipelines.ElasticNetRegressor method\), 1068](#)
[clone\(\) \(evalml.pipelines.Estimator method\), 1070](#)
[clone\(\) \(evalml.pipelines.ExtraTreesClassifier method\), 1073](#)
[clone\(\) \(evalml.pipelines.ExtraTreesRegressor method\), 1076](#)
[clone\(\) \(evalml.pipelines.FeatureSelector method\), 1078](#)
[clone\(\) \(evalml.pipelines.KNeighborsClassifier method\), 1081](#)
[clone\(\) \(evalml.pipelines.LightGBMClassifier method\), 1084](#)
[clone\(\) \(evalml.pipelines.LightGBMRegressor method\), 1086](#)
[clone\(\) \(evalml.pipelines.LinearRegressor method\), 1089](#)
[clone\(\) \(evalml.pipelines.LogisticRegressionClassifier method\), 1091](#)
[clone\(\) \(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassifier method\), 1002](#)
[clone\(\) \(evalml.pipelines.MulticlassClassificationPipeline method\), 1094](#)
[clone\(\) \(evalml.pipelines.OneHotEncoder method\), 1098](#)
[clone\(\) \(evalml.pipelines.PerColumnImputer method\), 1101](#)
[clone\(\) \(evalml.pipelines.pipeline_base.PipelineBase method\), 1106](#)
[clone\(\) \(evalml.pipelines.PipelineBase method\), 1103](#)
[clone\(\) \(evalml.pipelines.RandomForestClassifier method\), 1107](#)
[clone\(\) \(evalml.pipelines.RandomForestRegressor method\), 1109](#)
[clone\(\) \(evalml.pipelines.regression_pipeline.RegressionPipeline method\), 1012](#)
[clone\(\) \(evalml.pipelines.RegressionPipeline method\), 1112](#)
[clone\(\) \(evalml.pipelines.RFClassifierSelectFromModel method\), 1115](#)
[clone\(\) \(evalml.pipelines.RFRegressorSelectFromModel method\), 1118](#)
[clone\(\) \(evalml.pipelines.SimpleImputer method\), 1120](#)
[clone\(\) \(evalml.pipelines.StackedEnsembleClassifier method\), 1123](#)
[clone\(\) \(evalml.pipelines.StackedEnsembleRegressor method\), 1126](#)
[clone\(\) \(evalml.pipelines.StandardScaler method\), 1128](#)
[clone\(\) \(evalml.pipelines.SVMClassifier method\), 1130](#)
[clone\(\) \(evalml.pipelines.SVMRegressor method\), 1132](#)
[clone\(\) \(evalml.pipelines.TargetEncoder method\), 1135](#)
[clone\(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassifier method\), 1016](#)

`clone()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1021
`clone()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline` method), 1025
`clone()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` method), 1029
`clone()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` method), 1138
`clone()` (`evalml.pipelines.TimeSeriesClassificationPipeline` method), 1012
`clone()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` method), 1146
`clone()` (`evalml.pipelines.TimeSeriesRegressionPipeline` method), 1150
`clone()` (`evalml.pipelines.Transformer` method), 1154
`clone()` (`evalml.pipelines.XGBoostClassifier` method), 1156
`clone()` (`evalml.pipelines.XGBoostRegressor` method), 1158
`ColumnSelector` (class in `evalml.pipelines.components.transformers.column_selectors`), 775
`ComponentBase` (class in `evalml.pipelines.components`), 870
`ComponentBase` (class in `evalml.pipelines.components.component_base`), 848
`ComponentBaseMeta` (class in `evalml.pipelines.components`), 871
`ComponentBaseMeta` (class in `evalml.pipelines.components.component_base_meta`), 850
`ComponentGraph` (class in `evalml.pipelines`), 1051
`ComponentGraph` (class in `evalml.pipelines.component_graph`), 997
`ComponentNotYetFittedError`, 229, 230
`compute_estimator_features()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` method), 988
`compute_estimator_features()` (`evalml.pipelines.BinaryClassificationPipeline` method), 1039
`compute_estimator_features()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` method), 993
`compute_estimator_features()` (`evalml.pipelines.ClassificationPipeline` method), 1048
`compute_estimator_features()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` method), 1002
`compute_estimator_features()` (`evalml.pipelines.MulticlassClassificationPipeline` method), 1094
`compute_estimator_features()` (`evalml.pipelines.pipeline_base.PipelineBase` method), 1103
`compute_estimator_features()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` method), 1112
`compute_estimator_features()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1016
`compute_estimator_features()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline` method), 1106
`compute_estimator_features()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` method), 1021
`compute_estimator_features()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` method), 1150
`compute_estimator_features()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` method), 1138
`compute_estimator_features()` (`evalml.pipelines.TimeSeriesClassificationPipeline` method), 1142
`compute_estimator_features()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` method), 1146
`compute_estimator_features()` (`evalml.pipelines.TimeSeriesRegressionPipeline` method), 1150
`compute_estimator_features()` (`evalml.pipelines.ComponentGraph` method), 1052
`compute_estimator_features()` (`evalml.pipelines.ComponentGraph` method), 1052
`compute_order()` (`evalml.pipelines.component_graph.ComponentGraph` property), 998
`compute_order()` (`evalml.pipelines.ComponentGraph` property), 1052
`confusion_matrix()` (in module `evalml.model_understanding`), 256
`confusion_matrix()` (in module `evalml.model_understanding.graphs`), 243
`convert_to_seconds()` (in module `evalml.utils`), 1195
`convert_to_seconds()` (in module `evalml.utils.gen_utils`), 1190

<code>DataCheckAction</code> (class in <code>evalml.data_checks.data_check_action</code>), 186	in <code>decision_function()</code> (<code>evalml.objectives.AccuracyBinary</code> method), 351
<code>DataCheckActionCode</code> (class in <code>evalml.data_checks</code>), 211	in <code>decision_function()</code> (<code>evalml.objectives.AUC</code> method), 355
<code>DataCheckActionCode</code> (class in <code>evalml.data_checks.data_check_action_code</code>), 187	in <code>decision_function()</code> (<code>evalml.objectives.BalancedAccuracyBinary</code> method), 362
<code>DataCheckError</code> (class in <code>evalml.data_checks</code>), 211	in <code>decision_function()</code> (<code>evalml.objectives.binary_classification_objective.BinaryClassificationObjective</code> method), 266
<code>DataCheckError</code> (class in <code>evalml.data_checks.data_check_message</code>), 188	in <code>decision_function()</code> (<code>evalml.objectives.BinaryClassificationObjective</code> method), 365
<code>DataCheckInitError</code> , 229, 230	in <code>decision_function()</code> (<code>evalml.objectives.cost_benefit_matrix.CostBenefitMatrix</code> method), 269
<code>DataCheckMessage</code> (class in <code>evalml.data_checks</code>), 212	in <code>decision_function()</code> (<code>evalml.objectives.CostBenefitMatrix</code> method), 368
<code>DataCheckMessage</code> (class in <code>evalml.data_checks.data_check_message</code>), 188	in <code>decision_function()</code> (<code>evalml.objectives.F1</code> method), 372
<code>DataCheckMessageCode</code> (class in <code>evalml.data_checks</code>), 212	in <code>decision_function()</code> (<code>evalml.objectives.fraud_cost.FraudCost</code> method), 272
<code>DataCheckMessageCode</code> (class in <code>evalml.data_checks.data_check_message_code</code>), 189	in <code>decision_function()</code> (<code>evalml.objectives.FraudCost</code> method), 379
<code>DataCheckMessageType</code> (class in <code>evalml.data_checks</code>), 214	in <code>decision_function()</code> (<code>evalml.objectives.lead_scoring.LeadScoring</code> method), 274
<code>DataCheckMessageType</code> (class in <code>evalml.data_checks.data_check_message_type</code>), 191	in <code>decision_function()</code> (<code>evalml.objectives.LeadScoring</code> method), 382
<code>DataChecks</code> (class in <code>evalml.data_checks</code>), 214	in <code>decision_function()</code> (<code>evalml.objectives.LogLossBinary</code> method), 384
<code>DataChecks</code> (class in <code>evalml.data_checks.data_checks</code>), 192	in <code>decision_function()</code> (<code>evalml.objectives.MCCBinary</code> method), 393
<code>DataCheckWarning</code> (class in <code>evalml.data_checks</code>), 214	in <code>decision_function()</code> (<code>evalml.objectives.Precision</code> method), 405
<code>DataCheckWarning</code> (class in <code>evalml.data_checks.data_check_message</code>), 188	in <code>decision_function()</code> (<code>evalml.objectives.Recall</code> method), 414
<code>DateTimeFeaturizer</code> (class in <code>evalml.pipelines.components</code>), 872	in <code>decision_function()</code> (<code>evalml.objectives.sensitivity_low_alert.SensitivityLowAlert</code> method), 426
<code>DateTimeFeaturizer</code> (class in <code>evalml.pipelines.components.transformers</code>), 787	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AccuracyBinary</code> method), 287
<code>DateTimeFeaturizer</code> (class in <code>evalml.pipelines.components.transformers.preprocessing</code>), 726	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> method), 287
<code>DateTimeFeaturizer</code> (class in <code>evalml.pipelines.components.transformers.preprocessing.datetime_featurizer</code>), 705	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> method), 287
<code>DateTimeNaNDataCheck</code> (class in <code>evalml.data_checks</code>), 215	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> method), 287
<code>DateTimeNaNDataCheck</code> (class in <code>evalml.data_checks.datetime_nan_data_check</code>), 192	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> method), 287
<code>debug()</code> (<code>evalml.automl.engine.engine_base.JobLogger</code> method), 157	in <code>decision_function()</code> (<code>evalml.objectives.standard_metrics.AUC</code> method), 287

[method](#)), 291
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.BalancedAccuracyBinary](#) [method](#)), 298
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.F1](#) [method](#)), 303
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.LogLossBinary](#) [method](#)), 310
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.MCCBinary](#) [method](#)), 319
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.Precision](#) [method](#)), 327
[decision_function\(\)](#) ([evalml.objectives.standard_metrics.Recall](#) [method](#)), 336
[decision_tree_data_from_estimator\(\)](#) (in [module evalml.model_understanding.graphs](#)), 244
[decision_tree_data_from_pipeline\(\)](#) (in [module evalml.model_understanding.graphs](#)), 244
[DecisionTreeClassifier](#) (class in [evalml.pipelines](#)), 1054
[DecisionTreeClassifier](#) (class in [evalml.pipelines.components](#)), 874
[DecisionTreeClassifier](#) (class in [evalml.pipelines.components.estimators](#)), 593
[DecisionTreeClassifier](#) (class in [evalml.pipelines.components.estimators.classifiers](#)), 484
[DecisionTreeClassifier](#) (class in [evalml.pipelines.components.estimators.classifiers.decision_tree_classifier](#)), 451
[DecisionTreeRegressor](#) (class in [evalml.pipelines](#)), 1057
[DecisionTreeRegressor](#) (class in [evalml.pipelines.components](#)), 877
[DecisionTreeRegressor](#) (class in [evalml.pipelines.components.estimators](#)), 596
[DecisionTreeRegressor](#) (class in [evalml.pipelines.components.estimators.regressors](#)), 553
[DecisionTreeRegressor](#) (class in [evalml.pipelines.components.estimators.regressors.decision_tree_regressor](#)), 519
[DEFAULT_METRICS](#) (in [module evalml.model_understanding.prediction_explanations.explanations](#)), 235
[default_parameters\(\)](#) ([evalml.pipelines.ARIMARegressor](#) [method](#)), 1036
[default_parameters\(\)](#) ([evalml.pipelines.CatBoostClassifier](#) [method](#)), 1043
[default_parameters\(\)](#) ([evalml.pipelines.CatBoostRegressor](#) [method](#)), 1045
[default_parameters\(\)](#) ([evalml.pipelines.component_graph.ComponentGraph](#) [property](#)), 998
[default_parameters\(\)](#) ([evalml.pipelines.ComponentGraph](#) [property](#)), 1052
[default_parameters\(\)](#) ([evalml.pipelines.components.ARIMARegressor](#) [method](#)), 858
[default_parameters\(\)](#) ([evalml.pipelines.components.BaselineClassifier](#) [method](#)), 861
[default_parameters\(\)](#) ([evalml.pipelines.components.BaselineRegressor](#) [method](#)), 863
[default_parameters\(\)](#) ([evalml.pipelines.components.CatBoostClassifier](#) [method](#)), 866
[default_parameters\(\)](#) ([evalml.pipelines.components.CatBoostRegressor](#) [method](#)), 868
[default_parameters\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#) [method](#)), 849
[default_parameters\(\)](#) ([evalml.pipelines.components.ComponentBase](#) [method](#)), 870
[default_parameters\(\)](#) ([evalml.pipelines.components.DateFeatureTransformer](#) [method](#)), 873
[default_parameters\(\)](#) ([evalml.pipelines.components.DecisionTreeClassifier](#) [method](#)), 876
[default_parameters\(\)](#) ([evalml.pipelines.components.DecisionTreeRegressor](#) [method](#)), 879
[default_parameters\(\)](#) ([evalml.pipelines.components.DelayedFeatureTransformer](#) [method](#)), 881
[default_parameters\(\)](#) ([evalml.pipelines.components.DFSTransformer](#) [method](#)), 883
[default_parameters\(\)](#) ([evalml.pipelines.components.DropColumns](#) [method](#)), 885

<code>default_parameters()</code> (<code>evalml.pipelines.components.DropNullColumns</code> <code>method</code>), 887	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.CatBoostClassifier</code> <code>method</code>), 482
<code>default_parameters()</code> (<code>evalml.pipelines.components.ElasticNetClassifier</code> <code>method</code>), 890	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.decision_tree_classifier</code> <code>method</code>), 452
<code>default_parameters()</code> (<code>evalml.pipelines.components.ElasticNetRegressor</code> <code>method</code>), 893	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier</code> <code>method</code>), 485
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble_base</code> <code>method</code>), 429	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.elasticnet_classifier</code> <code>method</code>), 455
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble_classifier</code> <code>method</code>), 432	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier</code> <code>method</code>), 488
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble_regressor</code> <code>method</code>), 435	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.et_classifier</code> <code>method</code>), 459
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleBase</code> <code>method</code>), 438	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier</code> <code>method</code>), 491
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleClassifier</code> <code>method</code>), 441	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.kneighbors_classifier</code> <code>method</code>), 462
<code>default_parameters()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleRegressor</code> <code>method</code>), 443	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier</code> <code>method</code>), 494
<code>default_parameters()</code> (<code>evalml.pipelines.components.Estimator</code> <code>method</code>), 895	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.lightgbm_classifier</code> <code>method</code>), 465
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.ARIMAREgressor</code> <code>method</code>), 581	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.LightGBMClassifier</code> <code>method</code>), 497
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.BaselineClassifier</code> <code>method</code>), 584	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier</code> <code>method</code>), 468
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.BaselineRegressor</code> <code>method</code>), 586	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier</code> <code>method</code>), 500
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.CatBoostClassifier</code> <code>method</code>), 589	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.RandomForestClassifier</code> <code>method</code>), 503
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.CatBoostRegressor</code> <code>method</code>), 591	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.rf_classifier</code> <code>method</code>), 471
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.baseline_classifier</code> <code>method</code>), 446	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.svm_classifier</code> <code>method</code>), 474
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.BaselineClassifier</code> <code>method</code>), 480	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.SVMClassifier</code> <code>method</code>), 505
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.catboost_classifier</code> <code>method</code>), 449	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.classifiers.xgboost_classifier</code> <code>method</code>), 477

default_parameters() (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier, method), 508	default_parameters() (evalml.pipelines.components.estimators.regressors.baseline_regressor, method), 514
default_parameters() (evalml.pipelines.components.estimators.DecisionTreeClassifier, method), 594	default_parameters() (evalml.pipelines.components.estimators.regressors.BaselineRegressor, method), 549
default_parameters() (evalml.pipelines.components.estimators.DecisionTreeRegressor, method), 597	default_parameters() (evalml.pipelines.components.estimators.regressors.catboost_regressor, method), 517
default_parameters() (evalml.pipelines.components.estimators.ElasticNetClassifier, method), 600	default_parameters() (evalml.pipelines.components.estimators.regressors.CatBoostRegressor, method), 551
default_parameters() (evalml.pipelines.components.estimators.ElasticNetRegressor, method), 603	default_parameters() (evalml.pipelines.components.estimators.regressors.decision_tree_regressor, method), 520
default_parameters() (evalml.pipelines.components.estimators.Estimator, method), 605	default_parameters() (evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor, method), 554
default_parameters() (evalml.pipelines.components.estimators.estimator.Estimator, method), 578	default_parameters() (evalml.pipelines.components.estimators.regressors.elasticnet_regressor, method), 523
default_parameters() (evalml.pipelines.components.estimators.ExtraTreesClassifier, method), 608	default_parameters() (evalml.pipelines.components.estimators.regressors.ElasticNetRegressor, method), 557
default_parameters() (evalml.pipelines.components.estimators.ExtraTreesRegressor, method), 611	default_parameters() (evalml.pipelines.components.estimators.regressors.et_regressor.LinearRegressor, method), 526
default_parameters() (evalml.pipelines.components.estimators.KNeighborsClassifier, method), 614	default_parameters() (evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor, method), 560
default_parameters() (evalml.pipelines.components.estimators.LightGBMClassifier, method), 616	default_parameters() (evalml.pipelines.components.estimators.regressors.lightgbm_regressor, method), 529
default_parameters() (evalml.pipelines.components.estimators.LightGBMRegressor, method), 619	default_parameters() (evalml.pipelines.components.estimators.regressors.LightGBMRegressor, method), 563
default_parameters() (evalml.pipelines.components.estimators.LinearRegressor, method), 622	default_parameters() (evalml.pipelines.components.estimators.regressors.linear_regressor.LinearRegressor, method), 532
default_parameters() (evalml.pipelines.components.estimators.LogisticRegressionClassifier, method), 624	default_parameters() (evalml.pipelines.components.estimators.regressors.LinearRegressor, method), 565
default_parameters() (evalml.pipelines.components.estimators.RandomForestClassifier, method), 627	default_parameters() (evalml.pipelines.components.estimators.regressors.RandomForestRegressor, method), 568
default_parameters() (evalml.pipelines.components.estimators.RandomForestRegressor, method), 629	default_parameters() (evalml.pipelines.components.estimators.regressors.rf_regressor.LinearRegressor, method), 535
default_parameters() (evalml.pipelines.components.estimators.regressors.ard_regressor.ArimaRegressor, method), 511	default_parameters() (evalml.pipelines.components.estimators.regressors.svm_regressor.LinearRegressor, method), 538
default_parameters() (evalml.pipelines.components.estimators.regressors.ARIMARegressor, method), 546	default_parameters() (evalml.pipelines.components.estimators.regressors.SVMRegressor, method), 570

<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.time_series_baseline_pipeline_estimator</code> <code>method</code>), 540	<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.time_series_baseline_pipeline_estimator</code> <code>method</code>), 921
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.TimeSeriesBaselinePipelineEstimator</code> <code>method</code>), 572	<code>default_parameters()</code> (<code>evalml.pipelines.components.LSA</code> <code>method</code>), 923
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.xgboost_pipeline_estimator</code> <code>method</code>), 543	<code>default_parameters()</code> (<code>evalml.pipelines.components.XGBoostRegressor</code> <code>method</code>), 926
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.regressors.XGBoostRegressor</code> <code>method</code>), 575	<code>default_parameters()</code> (<code>evalml.pipelines.components.PCA</code> <code>method</code>), 928
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.SVMClassifier</code> <code>method</code>), 632	<code>default_parameters()</code> (<code>evalml.pipelines.components.PerColumnImputer</code> <code>method</code>), 931
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.SVMRegressor</code> <code>method</code>), 634	<code>default_parameters()</code> (<code>evalml.pipelines.components.PolynomialDetrender</code> <code>method</code>), 933
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator</code> <code>method</code>), 636	<code>default_parameters()</code> (<code>evalml.pipelines.components.RandomForestClassifier</code> <code>method</code>), 935
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.XGBoostClassifier</code> <code>method</code>), 639	<code>default_parameters()</code> (<code>evalml.pipelines.components.RandomForestRegressor</code> <code>method</code>), 938
<code>default_parameters()</code> (<code>evalml.pipelines.components.estimators.XGBoostRegressor</code> <code>method</code>), 641	<code>default_parameters()</code> (<code>evalml.pipelines.components.RFClassifierSelectFromModel</code> <code>method</code>), 940
<code>default_parameters()</code> (<code>evalml.pipelines.components.ExtraTreesClassifier</code> <code>method</code>), 898	<code>default_parameters()</code> (<code>evalml.pipelines.components.RFRegressorSelectFromModel</code> <code>method</code>), 943
<code>default_parameters()</code> (<code>evalml.pipelines.components.ExtraTreesRegressor</code> <code>method</code>), 901	<code>default_parameters()</code> (<code>evalml.pipelines.components.SelectColumns</code> <code>method</code>), 945
<code>default_parameters()</code> (<code>evalml.pipelines.components.FeatureSelector</code> <code>method</code>), 903	<code>default_parameters()</code> (<code>evalml.pipelines.components.SimpleImputer</code> <code>method</code>), 947
<code>default_parameters()</code> (<code>evalml.pipelines.components.Imputer</code> <code>method</code>), 905	<code>default_parameters()</code> (<code>evalml.pipelines.components.SMOTENCSampler</code> <code>method</code>), 950
<code>default_parameters()</code> (<code>evalml.pipelines.components.KNeighborsClassifier</code> <code>method</code>), 908	<code>default_parameters()</code> (<code>evalml.pipelines.components.SMOTENSampler</code> <code>method</code>), 952
<code>default_parameters()</code> (<code>evalml.pipelines.components.LightGBMClassifier</code> <code>method</code>), 911	<code>default_parameters()</code> (<code>evalml.pipelines.components.SMOTESampler</code> <code>method</code>), 954
<code>default_parameters()</code> (<code>evalml.pipelines.components.LightGBMRegressor</code> <code>method</code>), 914	<code>default_parameters()</code> (<code>evalml.pipelines.components.StackedEnsembleClassifier</code> <code>method</code>), 957
<code>default_parameters()</code> (<code>evalml.pipelines.components.LinearDiscriminantAnalysis</code> <code>method</code>), 916	<code>default_parameters()</code> (<code>evalml.pipelines.components.StackedEnsembleRegressor</code> <code>method</code>), 959
<code>default_parameters()</code> (<code>evalml.pipelines.components.LinearRegressor</code> <code>method</code>), 919	<code>default_parameters()</code> (<code>evalml.pipelines.components.StandardScaler</code> <code>method</code>), 961

<code>default_parameters()</code> (<i>evalml.pipelines.components.SVMClassifier</i> <i>method</i>), 964	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.DropNullColumns</i> <i>method</i>), 797
<code>default_parameters()</code> (<i>evalml.pipelines.components.SVMRegressor</i> <i>method</i>), 966	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.encoders.onehot_encoder</i> <i>method</i>), 654
<code>default_parameters()</code> (<i>evalml.pipelines.components.TargetEncoder</i> <i>method</i>), 968	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.encoders.OneHotEncoder</i> <i>method</i>), 661
<code>default_parameters()</code> (<i>evalml.pipelines.components.TargetImputer</i> <i>method</i>), 971	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.encoders.target_encoder</i> <i>method</i>), 658
<code>default_parameters()</code> (<i>evalml.pipelines.components.TextFeaturizer</i> <i>method</i>), 973	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.encoders.TargetEncoder</i> <i>method</i>), 664
<code>default_parameters()</code> (<i>evalml.pipelines.components.TimeSeriesBaselineEstimator</i> <i>method</i>), 975	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.feature_selector</i> <i>method</i>), 667
<code>default_parameters()</code> (<i>evalml.pipelines.components.Transformer</i> <i>method</i>), 977	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.FeatureSelector</i> <i>method</i>), 676
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.column_selector.ColumnSelector</i> <i>method</i>), 775	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.rf_classifier</i> <i>method</i>), 670
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.column_selector.DropColumns</i> <i>method</i>), 777	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.rf_classifier</i> <i>method</i>), 673
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.column_selector.SelectColumns</i> <i>method</i>), 779	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.RFClassifier</i> <i>method</i>), 679
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.DateTimeFeaturizer</i> <i>method</i>), 788	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.feature_selection.RFClassifier</i> <i>method</i>), 682
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.DelayedFeatureTransformer</i> <i>method</i>), 790	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.FeatureSelector</i> <i>method</i>), 799
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.DFSTransformer</i> <i>method</i>), 792	<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.Imputer</i> <i>method</i>), 801
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_base.DimensionalityReductionPipelineBase</i> <i>method</i>), 644	<code>default_parameters()</code> (<i>evalml.pipelines.components.imputers.Imputer</i> <i>method</i>), 696
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_base.DimensionalityReductionPipelineBase</i> <i>method</i>), 649	<code>default_parameters()</code> (<i>evalml.pipelines.components.imputers.imputer.Imputer</i> <i>method</i>), 685
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_base.DimensionalityReductionPipelineBase</i> <i>method</i>), 651	<code>default_parameters()</code> (<i>evalml.pipelines.components.imputers.per_column.Imputer</i> <i>method</i>), 688
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.dimensionality_reduction_pipeline_base.DimensionalityReductionPipelineBase</i> <i>method</i>), 646	<code>default_parameters()</code> (<i>evalml.pipelines.components.imputers.PerColumnImputer</i> <i>method</i>), 698
<code>default_parameters()</code> (<i>evalml.pipelines.components.transformers.DropColumns</i> <i>method</i>), 795	<code>default_parameters()</code> (<i>evalml.pipelines.components.imputers.simple_imputer.SimpleImputer</i> <i>method</i>), 690

default_parameters() (evalml.pipelines.components.transformers.imputers.SimpleImputer, method), 701	default_parameters() (evalml.pipelines.components.transformers.preprocessing.lsa.LSA, method), 716
default_parameters() (evalml.pipelines.components.transformers.imputers.target_encoder.TargetEncoder, method), 693	default_parameters() (evalml.pipelines.components.transformers.preprocessing.polynomial.Polynomial, method), 719
default_parameters() (evalml.pipelines.components.transformers.imputers.TargetImputer, method), 703	default_parameters() (evalml.pipelines.components.transformers.preprocessing.Polynomial, method), 738
default_parameters() (evalml.pipelines.components.transformers.LinearDiscriminantAnalysis, method), 804	default_parameters() (evalml.pipelines.components.transformers.preprocessing.text_featurizer.TextFeaturizer, method), 721
default_parameters() (evalml.pipelines.components.transformers.LSA, method), 806	default_parameters() (evalml.pipelines.components.transformers.preprocessing.text_transformer.TextTransformer, method), 724
default_parameters() (evalml.pipelines.components.transformers.OneHotEncoder, method), 808	default_parameters() (evalml.pipelines.components.transformers.preprocessing.TextFeaturizer, method), 741
default_parameters() (evalml.pipelines.components.transformers.PCA, method), 811	default_parameters() (evalml.pipelines.components.transformers.preprocessing.TextTransformer, method), 743
default_parameters() (evalml.pipelines.components.transformers.PerColumnImputer, method), 813	default_parameters() (evalml.pipelines.components.transformers.RFClassifierSelectFromModel, method), 818
default_parameters() (evalml.pipelines.components.transformers.PolynomialDetrender, method), 815	default_parameters() (evalml.pipelines.components.transformers.RFRegressorSelectFromModel, method), 821
default_parameters() (evalml.pipelines.components.transformers.preprocessing.datetime_datetime.DateTimeFeaturizer, method), 706	default_parameters() (evalml.pipelines.components.transformers.samplers.base_sampler.BaseSampler, method), 746
default_parameters() (evalml.pipelines.components.transformers.preprocessing.DateFeaturizer, method), 727	default_parameters() (evalml.pipelines.components.transformers.samplers.base_sampler.BaseSampler, method), 748
default_parameters() (evalml.pipelines.components.transformers.preprocessing.datetime_datetime.DateTimeFeaturizer, method), 709	default_parameters() (evalml.pipelines.components.transformers.samplers.base_sampler.BaseSampler, method), 751
default_parameters() (evalml.pipelines.components.transformers.preprocessing.DateFeaturizer, method), 730	default_parameters() (evalml.pipelines.components.transformers.samplers.oversampler.Oversampler, method), 753
default_parameters() (evalml.pipelines.components.transformers.preprocessing.DatetimeFeaturizer, method), 732	default_parameters() (evalml.pipelines.components.transformers.samplers.oversampler.Oversampler, method), 755
default_parameters() (evalml.pipelines.components.transformers.preprocessing.datetime_datetime.DateTimeFeaturizer, method), 711	default_parameters() (evalml.pipelines.components.transformers.samplers.SMOTEKMeans, method), 761
default_parameters() (evalml.pipelines.components.transformers.preprocessing.DateFeaturizer, method), 734	default_parameters() (evalml.pipelines.components.transformers.samplers.SMOTENSA, method), 763
default_parameters() (evalml.pipelines.components.transformers.preprocessing.datetime_datetime.DateTimeFeaturizer, method), 714	default_parameters() (evalml.pipelines.components.transformers.samplers.SMOTESampling, method), 766
default_parameters() (evalml.pipelines.components.transformers.preprocessing.LSA, method), 736	default_parameters() (evalml.pipelines.components.transformers.samplers.Undersampler, method), 768

default_parameters() (evalml.pipelines.components.transformers.samplers.undersampler method), 758	default_parameters() (evalml.pipelines.components.XGBoostRegressor method), 985
default_parameters() (evalml.pipelines.components.transformers.scalers.standard_scaler method), 771	default_parameters() (evalml.pipelines.DecisionTreeClassifier method), 1055
default_parameters() (evalml.pipelines.components.transformers.scalers.StandardScaler method), 773	default_parameters() (evalml.pipelines.DecisionTreeRegressor method), 1058
default_parameters() (evalml.pipelines.components.transformers.SelectColumns method), 823	default_parameters() (evalml.pipelines.DelayedFeatureTransformer method), 1061
default_parameters() (evalml.pipelines.components.transformers.SimpleImputer method), 825	default_parameters() (evalml.pipelines.DFSTransformer method), 1063
default_parameters() (evalml.pipelines.components.transformers.SMOTENCSampler method), 828	default_parameters() (evalml.pipelines.ElasticNetClassifier method), 1066
default_parameters() (evalml.pipelines.components.transformers.SMOTENSampler method), 830	default_parameters() (evalml.pipelines.ElasticNetRegressor method), 1068
default_parameters() (evalml.pipelines.components.transformers.SMOTESampler method), 832	default_parameters() (evalml.pipelines.Estimator method), 1070
default_parameters() (evalml.pipelines.components.transformers.StandardScaler method), 834	default_parameters() (evalml.pipelines.ExtraTreesClassifier method), 1073
default_parameters() (evalml.pipelines.components.transformers.TargetEncoder method), 837	default_parameters() (evalml.pipelines.ExtraTreesRegressor method), 1076
default_parameters() (evalml.pipelines.components.transformers.TargetImputer method), 839	default_parameters() (evalml.pipelines.FeatureSelector method), 1078
default_parameters() (evalml.pipelines.components.transformers.TextFeaturizer method), 841	default_parameters() (evalml.pipelines.KNeighborsClassifier method), 1081
default_parameters() (evalml.pipelines.components.transformers.Transformer method), 843	default_parameters() (evalml.pipelines.LightGBMClassifier method), 1084
default_parameters() (evalml.pipelines.components.transformers.transformer.TargetEncoder method), 782	default_parameters() (evalml.pipelines.LightGBMRegressor method), 1086
default_parameters() (evalml.pipelines.components.transformers.transformer.TargetEncoder method), 784	default_parameters() (evalml.pipelines.LinearRegressor method), 1089
default_parameters() (evalml.pipelines.components.transformers.Undersampler method), 846	default_parameters() (evalml.pipelines.LogisticRegressionClassifier method), 1091
default_parameters() (evalml.pipelines.components.Undersampler method), 980	default_parameters() (evalml.pipelines.OneHotEncoder method), 1098
default_parameters() (evalml.pipelines.components.XGBoostClassifier method), 982	default_parameters() (evalml.pipelines.PerColumnImputer method), 1101
	default_parameters()

(*evalml.pipelines.RandomForestClassifier* method), 1107

default_parameters() (*evalml.pipelines.RandomForestRegressor* method), 1109

default_parameters() (*evalml.pipelines.RFClassifierSelectFromModel* method), 1115

default_parameters() (*evalml.pipelines.RFRegressorSelectFromModel* method), 1118

default_parameters() (*evalml.pipelines.SimpleImputer* method), 1120

default_parameters() (*evalml.pipelines.StackedEnsembleClassifier* method), 1123

default_parameters() (*evalml.pipelines.StackedEnsembleRegressor* method), 1126

default_parameters() (*evalml.pipelines.StandardScaler* method), 1128

default_parameters() (*evalml.pipelines.SVMClassifier* method), 1130

default_parameters() (*evalml.pipelines.SVMRegressor* method), 1132

default_parameters() (*evalml.pipelines.TargetEncoder* method), 1135

default_parameters() (*evalml.pipelines.Transformer* method), 1154

default_parameters() (*evalml.pipelines.XGBoostClassifier* method), 1156

default_parameters() (*evalml.pipelines.XGBoostRegressor* method), 1159

DefaultDataChecks (class in *evalml.data_checks*), 215

DefaultDataChecks (class in *evalml.data_checks.default_data_checks*), 194

DelayedFeatureTransformer (class in *evalml.pipelines*), 1060

DelayedFeatureTransformer (class in *evalml.pipelines.components*), 880

DelayedFeatureTransformer (class in *evalml.pipelines.components.transformers*), 789

DelayedFeatureTransformer (class in *evalml.pipelines.components.transformers.preprocessing*), 729

DelayedFeatureTransformer (class in *evalml.pipelines.components.transformers.preprocessing.delayed*), 708

deprecate_arg() (in module *evalml.utils*), 1195

deprecate_arg() (in module *evalml.utils.gen_utils*), 1190

describe() (*evalml.pipelines.ARIMARegressor* method), 1036

describe() (*evalml.pipelines.binary_classification_pipeline.BinaryClassifier* method), 988

describe() (*evalml.pipelines.BinaryClassificationPipeline* method), 1039

describe() (*evalml.pipelines.CatBoostClassifier* method), 1043

describe() (*evalml.pipelines.CatBoostRegressor* method), 1045

describe() (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 994

describe() (*evalml.pipelines.ClassificationPipeline* method), 1048

describe() (*evalml.pipelines.component_graph.ComponentGraph* method), 998

describe() (*evalml.pipelines.ComponentGraph* method), 1052

describe() (*evalml.pipelines.components.ARIMARegressor* method), 859

describe() (*evalml.pipelines.components.BaselineClassifier* method), 861

describe() (*evalml.pipelines.components.BaselineRegressor* method), 863

describe() (*evalml.pipelines.components.CatBoostClassifier* method), 866

describe() (*evalml.pipelines.components.CatBoostRegressor* method), 868

describe() (*evalml.pipelines.components.component_base.ComponentBase* method), 849

describe() (*evalml.pipelines.components.ComponentBase* method), 870

describe() (*evalml.pipelines.components.DateTimeFeaturizer* method), 873

describe() (*evalml.pipelines.components.DecisionTreeClassifier* method), 876

describe() (*evalml.pipelines.components.DecisionTreeRegressor* method), 879

describe() (*evalml.pipelines.components.DelayedFeatureTransformer* method), 881

describe() (*evalml.pipelines.components.DFSTransformer* method), 883

describe() (*evalml.pipelines.components.DropColumns* method), 885

describe() (*evalml.pipelines.components.DropNullColumns* method), 887

describe() (*evalml.pipelines.components.ElasticNetClassifier*

method), 890

describe() (evalml.pipelines.components.ElasticNetRegressor method), 893

describe() (evalml.pipelines.components.ensemble.stacked_ensemble method), 429

describe() (evalml.pipelines.components.ensemble.stacked_ensemble method), 432

describe() (evalml.pipelines.components.ensemble.stacked_ensemble method), 435

describe() (evalml.pipelines.components.ensemble.StackedEnsembleBase method), 438

describe() (evalml.pipelines.components.ensemble.StackedEnsembleClassifier method), 441

describe() (evalml.pipelines.components.ensemble.StackedEnsembleRegressor method), 443

describe() (evalml.pipelines.components.Estimator method), 895

describe() (evalml.pipelines.components.estimators.ARIMARegressor method), 582

describe() (evalml.pipelines.components.estimators.BaselineClassifier method), 584

describe() (evalml.pipelines.components.estimators.BaselineRegressor method), 586

describe() (evalml.pipelines.components.estimators.CatBoostClassifier method), 589

describe() (evalml.pipelines.components.estimators.CatBoostRegressor method), 592

describe() (evalml.pipelines.components.estimators.classifiers.baseline_classifier method), 446

describe() (evalml.pipelines.components.estimators.classifiers.BaselineClassifier method), 480

describe() (evalml.pipelines.components.estimators.classifiers.catboost_classifier method), 449

describe() (evalml.pipelines.components.estimators.classifiers.CatBoostClassifier method), 482

describe() (evalml.pipelines.components.estimators.classifiers.decision_tree_classifier method), 452

describe() (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier method), 485

describe() (evalml.pipelines.components.estimators.classifiers.elasticnet_classifier method), 456

describe() (evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier method), 488

describe() (evalml.pipelines.components.estimators.classifiers.extra_trees_classifier method), 459

describe() (evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier method), 491

describe() (evalml.pipelines.components.estimators.classifiers.kneighbors_classifier method), 462

describe() (evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier method), 494

describe() (evalml.pipelines.components.estimators.classifiers.lightgbm_classifier method), 465

describe() (evalml.pipelines.components.estimators.classifiers.LightGBMClassifier method), 497

describe() (evalml.pipelines.components.estimators.classifiers.logistic_regression method), 468

describe() (evalml.pipelines.components.estimators.classifiers.LogisticRegression method), 500

describe() (evalml.pipelines.components.estimators.classifiers.RandomForestClassifier method), 503

describe() (evalml.pipelines.components.estimators.classifiers.rf_classifier method), 471

describe() (evalml.pipelines.components.estimators.classifiers.svm_classifier method), 474

describe() (evalml.pipelines.components.estimators.classifiers.SVMClassifier method), 505

describe() (evalml.pipelines.components.estimators.classifiers.xgboost_classifier method), 477

describe() (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier method), 508

describe() (evalml.pipelines.components.estimators.DecisionTreeRegressor method), 594

describe() (evalml.pipelines.components.estimators.DecisionTreeRegressor method), 597

describe() (evalml.pipelines.components.estimators.ElasticNetClassifier method), 600

describe() (evalml.pipelines.components.estimators.ElasticNetRegressor method), 603

describe() (evalml.pipelines.components.estimators.Estimator method), 605

describe() (evalml.pipelines.components.estimators.extra_trees_classifier method), 578

describe() (evalml.pipelines.components.estimators.ExtraTreesClassifier method), 608

describe() (evalml.pipelines.components.estimators.extra_trees_classifier method), 611

describe() (evalml.pipelines.components.estimators.KNeighborsClassifier method), 614

describe() (evalml.pipelines.components.estimators.LightGBMClassifier method), 617

describe() (evalml.pipelines.components.estimators.LightGBMRegressor method), 619

describe() (evalml.pipelines.components.estimators.LinearRegressor method), 622

describe() (evalml.pipelines.components.estimators.LogisticRegression method), 624

describe() (evalml.pipelines.components.estimators.RandomForestClassifier method), 627

describe() (evalml.pipelines.components.estimators.RandomForestRegressor method), 629

describe() (evalml.pipelines.components.estimators.regressors.arima_regressor method), 511

describe() (evalml.pipelines.components.estimators.regressors.ARIMARegressor method), 547

describe() (evalml.pipelines.components.estimators.regressors.baseline_regressor method), 514

describe() (evalml.pipelines.components.estimators.regressors.BaselineRegressor method), 549

method), 549

describe() (evalml.pipelines.components.estimators.regressor.classic.boosting.CatBoostRegressor), 517

describe() (evalml.pipelines.components.estimators.regressor.classic.boosting.GradientBoostingRegressor), 552

describe() (evalml.pipelines.components.estimators.regressor.classic.decision_tree.DecisionTreeRegressor), 520

describe() (evalml.pipelines.components.estimators.regressor.classic.decision_tree.DecisionTreeRegressor), 554

describe() (evalml.pipelines.components.estimators.regressor.classic.elastic_net.ElasticNetRegressor), 523

describe() (evalml.pipelines.components.estimators.regressor.classic.elastic_net.ElasticNetRegressor), 557

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 526

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 560

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 529

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 563

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 532

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 565

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 568

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 535

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 538

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 570

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 540

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 573

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 543

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 575

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 632

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 634

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 636

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 639

describe() (evalml.pipelines.components.estimators.regressor.classic.linear_regression.LinearRegression), 641

describe() (evalml.pipelines.components.ExtraTreesClassifier), 898

describe() (evalml.pipelines.components.ExtraTreesRegressor), 901

describe() (evalml.pipelines.components.FeatureSelector), 903

describe() (evalml.pipelines.components.Imputer), 905

describe() (evalml.pipelines.components.LightGBMClassifier), 908

describe() (evalml.pipelines.components.LightGBMRegressor), 911

describe() (evalml.pipelines.components.LinearDiscriminantAnalysis), 914

describe() (evalml.pipelines.components.LinearRegression), 916

describe() (evalml.pipelines.components.LogisticRegressionClassifier), 919

describe() (evalml.pipelines.components.LSA), 922

describe() (evalml.pipelines.components.OneHotEncoder), 926

describe() (evalml.pipelines.components.PCA), 929

describe() (evalml.pipelines.components.PerColumnImputer), 931

describe() (evalml.pipelines.components.PolynomialDetrender), 933

describe() (evalml.pipelines.components.RandomForestClassifier), 936

describe() (evalml.pipelines.components.RandomForestRegressor), 938

describe() (evalml.pipelines.components.RFClassifierSelectFromModel), 940

describe() (evalml.pipelines.components.SelectColumns), 943

describe() (evalml.pipelines.components.SelectFromModel), 945

describe() (evalml.pipelines.components.SimpleImputer), 948

describe() (evalml.pipelines.components.SMOTE sampler), 950

describe() (evalml.pipelines.components.SMOTE sampler), 952

describe() (evalml.pipelines.components.SMOTE sampler), 954

describe() (evalml.pipelines.components.StackedEnsembleClassifier), 957

describe() (evalml.pipelines.components.StackedEnsembleRegressor), 959

describe() (evalml.pipelines.components.StandardScaler), 961

describe() (evalml.pipelines.components.SVMClassifier), 964

describe() (evalml.pipelines.components.SVMRegressor), 964

method), 966

describe() (evalml.pipelines.components.TargetEncoder, method), 968

describe() (evalml.pipelines.components.TargetImputer, method), 971

describe() (evalml.pipelines.components.TextFeaturizer, method), 973

describe() (evalml.pipelines.components.TimeSeriesBaselineEstimator, method), 975

describe() (evalml.pipelines.components.Transformer, method), 977

describe() (evalml.pipelines.components.transformers.ColumnSelector, method), 775

describe() (evalml.pipelines.components.transformers.ColumnSelector, method), 778

describe() (evalml.pipelines.components.transformers.ColumnSelector, method), 780

describe() (evalml.pipelines.components.transformers.DateTimeFeaturizer, method), 788

describe() (evalml.pipelines.components.transformers.DelayedFeatureTransformer, method), 790

describe() (evalml.pipelines.components.transformers.DESTransformer, method), 793

describe() (evalml.pipelines.components.transformers.dimensionality_reduction.Transformers.LSA, method), 644

describe() (evalml.pipelines.components.transformers.dimensionality_reduction.Transformers.LSA, method), 649

describe() (evalml.pipelines.components.transformers.dimensionality_reduction.Transformers.PCA, method), 651

describe() (evalml.pipelines.components.transformers.dimensionality_reduction.Transformers.PCA, method), 646

describe() (evalml.pipelines.components.transformers.DropColumn, method), 795

describe() (evalml.pipelines.components.transformers.DropNullColumns, method), 797

describe() (evalml.pipelines.components.transformers.ensemble.OneHotEmbedder, method), 654

describe() (evalml.pipelines.components.transformers.ensemble.OneHotEmbedder, method), 661

describe() (evalml.pipelines.components.transformers.ensemble.TargetEncoder, method), 658

describe() (evalml.pipelines.components.transformers.ensemble.TargetEncoder, method), 664

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 667

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 676

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 670

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 673

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 679

describe() (evalml.pipelines.components.transformers.feature_selection.FeatureSelector, method), 682

describe() (evalml.pipelines.components.transformers.FeatureSelector, method), 799

describe() (evalml.pipelines.components.transformers.Imputer, method), 801

describe() (evalml.pipelines.components.transformers.imputers.Imputer, method), 696

describe() (evalml.pipelines.components.transformers.imputers.impute, method), 685

describe() (evalml.pipelines.components.transformers.imputers.per_column, method), 688

describe() (evalml.pipelines.components.transformers.imputers.PerColumn, method), 699

describe() (evalml.pipelines.components.transformers.imputers.simple, method), 690

describe() (evalml.pipelines.components.transformers.imputers.Simple, method), 701

describe() (evalml.pipelines.components.transformers.imputers.target, method), 693

describe() (evalml.pipelines.components.transformers.imputers.Target, method), 703

describe() (evalml.pipelines.components.transformers.LinearDiscrimin, method), 804

describe() (evalml.pipelines.components.transformers.LinearDiscrimin, method), 806

describe() (evalml.pipelines.components.transformers.LinearDiscrimin, method), 808

describe() (evalml.pipelines.components.transformers.PCA, method), 811

describe() (evalml.pipelines.components.transformers.PerColumnImpu, method), 813

describe() (evalml.pipelines.components.transformers.PolynomialDete, method), 815

describe() (evalml.pipelines.components.transformers.preprocessing.d, method), 706

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 727

describe() (evalml.pipelines.components.transformers.preprocessing.d, method), 709

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 730

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 732

describe() (evalml.pipelines.components.transformers.preprocessing.d, method), 712

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 734

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 714

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 736

describe() (evalml.pipelines.components.transformers.preprocessing.D, method), 716

describe() (evalml.pipelines.components.transformers.preprocessing.p, method), 716

method), 719

describe() (evalml.pipelines.components.transformers.preprocessing.PolyomialPipelineComponent), 738

describe() (evalml.pipelines.components.transformers.preprocessing.TextFeaturizer), 722

describe() (evalml.pipelines.components.transformers.preprocessing.TextTransformer), 724

describe() (evalml.pipelines.components.transformers.preprocessing.TextTransformer), 741

describe() (evalml.pipelines.components.transformers.preprocessing.TextTransformer), 743

describe() (evalml.pipelines.components.transformers.RFClassifier), 818

describe() (evalml.pipelines.components.transformers.RFRegressor), 821

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 746

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 748

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 751

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 753

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 756

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 761

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 763

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 766

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 768

describe() (evalml.pipelines.components.transformers.samplers.base.Sampler), 758

describe() (evalml.pipelines.components.transformers.scalers.standard_scaler.StandardScaler), 771

describe() (evalml.pipelines.components.transformers.scalers.standard_scaler.StandardScaler), 773

describe() (evalml.pipelines.components.transformers.SelectColumns), 823

describe() (evalml.pipelines.components.transformers.SimpleImputer), 825

describe() (evalml.pipelines.components.transformers.SMOTECSampler), 828

describe() (evalml.pipelines.components.transformers.SMOTENSampler), 830

describe() (evalml.pipelines.components.transformers.SMOTESampler), 832

describe() (evalml.pipelines.components.transformers.StandardScaler), 834

describe() (evalml.pipelines.components.transformers.TargetEncoder), 837

describe() (evalml.pipelines.components.transformers.TargetImputer), 839

describe() (evalml.pipelines.components.transformers.TextFeaturizer), 841

describe() (evalml.pipelines.components.transformers.TextTransformer), 844

describe() (evalml.pipelines.components.transformers.transformer.TargetEncoder), 782

describe() (evalml.pipelines.components.transformers.transformer.TargetEncoder), 784

describe() (evalml.pipelines.components.transformers.Undersampler), 846

describe() (evalml.pipelines.components.Undersampler), 980

describe() (evalml.pipelines.components.XGBoostClassifier), 982

describe() (evalml.pipelines.components.XGBoostRegressor), 985

describe() (evalml.pipelines.components.XGBoostRegressor), 1056

describe() (evalml.pipelines.components.XGBoostRegressor), 1058

describe() (evalml.pipelines.components.XGBoostRegressor), 1061

describe() (evalml.pipelines.components.XGBoostRegressor), 1063

describe() (evalml.pipelines.components.XGBoostRegressor), 1066

describe() (evalml.pipelines.components.XGBoostRegressor), 1068

describe() (evalml.pipelines.components.XGBoostRegressor), 1070

describe() (evalml.pipelines.ExtraTreesClassifier), 1073

describe() (evalml.pipelines.ExtraTreesRegressor), 1076

describe() (evalml.pipelines.FeatureSelector), 1078

describe() (evalml.pipelines.KNeighborsClassifier), 1081

describe() (evalml.pipelines.LightGBMClassifier), 1084

describe() (evalml.pipelines.LightGBMRegressor), 1087

describe() (evalml.pipelines.LinearRegressor), 1089

describe() (evalml.pipelines.LogisticRegressionClassifier), 1092

describe() (evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline), 1002

describe() (evalml.pipelines.MulticlassClassificationPipeline), 1095

describe() (evalml.pipelines.OneHotEncoder), 1098

describe() (evalml.pipelines.PerColumnImputer), 1098

method), 1101

describe() (evalml.pipelines.pipeline_base.PipelineBase method), 1006

describe() (evalml.pipelines.PipelineBase method), 1104

describe() (evalml.pipelines.RandomForestClassifier method), 1107

describe() (evalml.pipelines.RandomForestRegressor method), 1109

describe() (evalml.pipelines.regression_pipeline.RegressionPipeline method), 1012

describe() (evalml.pipelines.RegressionPipeline method), 1112

describe() (evalml.pipelines.RFClassifierSelectFromModel method), 1116

describe() (evalml.pipelines.RFRegressorSelectFromModel method), 1118

describe() (evalml.pipelines.SimpleImputer method), 1121

describe() (evalml.pipelines.StackedEnsembleClassifier method), 1123

describe() (evalml.pipelines.StackedEnsembleRegressor method), 1126

describe() (evalml.pipelines.StandardScaler method), 1128

describe() (evalml.pipelines.SVMClassifier method), 1130

describe() (evalml.pipelines.SVMRegressor method), 1133

describe() (evalml.pipelines.TargetEncoder method), 1135

describe() (evalml.pipelines.time_series_classification_pipeline.EnableScrubbingClassifierPipeline method), 1016

describe() (evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method), 1021

describe() (evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline method), 1025

describe() (evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method), 1030

describe() (evalml.pipelines.TimeSeriesBinaryClassificationPipeline method), 1138

describe() (evalml.pipelines.TimeSeriesClassificationPipeline method), 1142

describe() (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method), 1147

describe() (evalml.pipelines.TimeSeriesRegressionPipeline method), 1151

describe() (evalml.pipelines.Transformer method), 1154

describe() (evalml.pipelines.XGBoostClassifier method), 1156

describe() (evalml.pipelines.XGBoostRegressor method), 1159

describe_pipeline() (evalml.automl.automl_search.AutoMLSearch method), 169

describe_pipeline() (evalml.automl.AutoMLSearch method), 179

describe_pipeline() (evalml.AutoMLSearch method), 1201

detect_problem_type() (in module evalml.problem_types), 1175

detect_problem_type() (in module evalml.problem_types.utils), 1173

DFSTransformer (class in evalml.pipelines), 1062

DFSTransformer (class in evalml.pipelines.components), 882

DFSTransformer (class in evalml.pipelines.components.transformers), 792

DFSTransformer (class in evalml.pipelines.components.transformers.preprocessing), 731

DFSTransformer (class in evalml.pipelines.components.transformers.preprocessing.feature_transformers), 713

done() (evalml.automl.engine.cf_engine.CFComputation method), 151

done() (evalml.automl.engine.dask_engine.DaskComputation method), 153

done() (evalml.automl.engine.engine_base.EngineComputation method), 156

done() (evalml.automl.engine.EngineComputation method), 163

done() (evalml.automl.engine.SequentialComputation method), 159

enable_scrubbing() (in module evalml.preprocessing), 1169

enable_scrubbing() (in module evalml.preprocessing.utils), 1167

enable_time_series_regression() (in module evalml.utils), 1195

fit_with_nans() (in module evalml.utils.gen_utils), 1190

DropColumns (class in evalml.pipelines.components), 885

DropColumns (class in evalml.pipelines.components.transformers), 794

DropColumns (class in evalml.pipelines.components.transformers.column_selectors), 777

DropNullColumns (class in evalml.pipelines.components), 887

DropNullColumns (class in evalml.pipelines.components.transformers), 796

DropNullColumns (class in Estimator (class in *evalml.pipelines*), 1069
evalml.pipelines.components.transformers.preprocessing.dropnull_columns), 894
733 Estimator (class in

DropNullColumns (class in *evalml.pipelines.components.estimators*),
evalml.pipelines.components.transformers.preprocessing.dropnull_columns),
711 Estimator (class in
evalml.pipelines.components.estimators.estimator),
577

E

ElasticNetClassifier (class in *evalml.pipelines*), *evalml*
1064 module, 143

ElasticNetClassifier (class in *evalml.pipelines.components*), 889 *evalml.automl*
module, 143

ElasticNetClassifier (class in *evalml.pipelines.components.estimators*),
598 *evalml.automl.automl_algorithm*
module, 143
evalml.automl.automl_algorithm.automl_algorithm

ElasticNetClassifier (class in *evalml.pipelines.components.estimators.classifiers*),
486 *evalml.automl.automl_algorithm.iterative_algorithm*
module, 145

ElasticNetClassifier (class in *evalml.pipelines.components.estimators.classifiers.elasticnet_classifier*),
454 *evalml.automl.automl_search*
module, 166

ElasticNetRegressor (class in *evalml.pipelines*),
1067 *evalml.automl.callbacks*
module, 171
evalml.automl.engine

ElasticNetRegressor (class in *evalml.pipelines.components*), 891 *evalml.automl.engine.cf_engine*

ElasticNetRegressor (class in *evalml.pipelines.components.estimators*),
601 *evalml.automl.engine.dask_engine*
module, 153

ElasticNetRegressor (class in *evalml.pipelines.components.estimators.regressors*),
556 *evalml.automl.engine.engine_base*
module, 155
evalml.automl.engine.sequential_engine

ElasticNetRegressor (class in *evalml.pipelines.components.estimators.regressors.elasticnet_regressor*),
522 *evalml.automl.pipeline_search_plots*
module, 172

EmptyDataChecks (class in *evalml.data_checks*), 216 *evalml.automl.utils*

EmptyDataChecks (class in *evalml.data_checks.utils*), 208 module, 173

EngineBase (class in *evalml.automl*), 181 *evalml.data_checks*
module, 183

EngineBase (class in *evalml.automl.engine*), 162 *evalml.data_checks.class_imbalance_data_check*

EngineBase (class in *evalml.automl.engine.engine_base*), 156 module, 183

EngineComputation (class in *evalml.automl.engine*), 163 *evalml.data_checks.data_check*
module, 185

EngineComputation (class in *evalml.automl.engine.engine_base*), 156 *evalml.data_checks.data_check_action*
module, 186

EnsembleMissingPipelinesError, 229, 230 *evalml.data_checks.data_check_action_code*
module, 187

error() (*evalml.automl.engine.engine_base.JobLogger*
method), 157 *evalml.data_checks.data_check_message*
module, 187

error_contains_nan (in module *evalml.data_checks.datetime_nan_data_check*),
193 *evalml.data_checks.data_check_message_code*
module, 189

error_contains_nan (in module *evalml.data_checks.natural_language_nan_data_check*),
200 *evalml.data_checks.data_check_message_type*
module, 191

evalml.data_checks.data_checks
module, 191

evalml.data_checks.datetime_nan_data_check	evalml.model_understanding.permutation_importance
module, 192	module, 252
evalml.data_checks.default_data_checks	evalml.model_understanding.prediction_explanations
module, 193	module, 234
evalml.data_checks.highly_null_data_check	evalml.model_understanding.prediction_explanations
module, 195	module, 234
evalml.data_checks.id_columns_data_check	evalml.objectives
module, 196	module, 265
evalml.data_checks.invalid_targets_data_check	evalml.objectives.binary_classification_objective
module, 197	module, 265
evalml.data_checks.multicollinearity_data_check	evalml.objectives.cost_benefit_matrix
module, 199	module, 268
evalml.data_checks.natural_language_nan_data_check	evalml.objectives.fraud_cost
module, 199	module, 270
evalml.data_checks.no_variance_data_check	evalml.objectives.lead_scoring
module, 201	module, 273
evalml.data_checks.outliers_data_check	evalml.objectives.multiclass_classification_objective
module, 202	module, 276
evalml.data_checks.sparsity_data_check	evalml.objectives.objective_base
module, 203	module, 278
evalml.data_checks.target_leakage_data_check	evalml.objectives.regression_objective
module, 204	module, 280
evalml.data_checks.uniqueness_data_check	evalml.objectives.sensitivity_low_alert
module, 206	module, 283
evalml.data_checks.utils	evalml.objectives.standard_metrics
module, 207	module, 285
evalml.demos	evalml.objectives.time_series_regression_objective
module, 226	module, 345
evalml.demos.breast_cancer	evalml.objectives.utils
module, 226	module, 347
evalml.demos.churn	evalml.pipelines
module, 226	module, 427
evalml.demos.diabetes	evalml.pipelines.binary_classification_pipeline
module, 227	module, 986
evalml.demos.fraud	evalml.pipelines.binary_classification_pipeline_mixin
module, 227	module, 991
evalml.demos.wine	evalml.pipelines.classification_pipeline
module, 228	module, 992
evalml.exceptions	evalml.pipelines.component_graph
module, 229	module, 996
evalml.exceptions.exceptions	evalml.pipelines.components
module, 229	module, 427
evalml.model_family	evalml.pipelines.components.component_base
module, 231	module, 848
evalml.model_family.model_family	evalml.pipelines.components.component_base_meta
module, 231	module, 850
evalml.model_family.utils	evalml.pipelines.components.ensemble
module, 232	module, 427
evalml.model_understanding	evalml.pipelines.components.ensemble.stacked_ensemble
module, 234	module, 427
evalml.model_understanding.force_plots	evalml.pipelines.components.ensemble.stacked_ensemble
module, 241	module, 430
evalml.model_understanding.graphs	evalml.pipelines.components.ensemble.stacked_ensemble
module, 242	module, 433

<code>evalml.pipelines.components.estimators</code>	<code>evalml.pipelines.components.transformers</code>
module, 445	module, 643
<code>evalml.pipelines.components.estimators.column_selector</code>	<code>evalml.pipelines.components.transformers.column_selector</code>
module, 445	module, 774
<code>evalml.pipelines.components.estimators.column_selector_base</code>	<code>evalml.pipelines.components.transformers.dimensionality_reduction</code>
module, 445	module, 643
<code>evalml.pipelines.components.estimators.column_selector_base.compose</code>	<code>evalml.pipelines.components.transformers.dimensionality_reduction.compose</code>
module, 448	module, 643
<code>evalml.pipelines.components.estimators.column_selector_base.transformers</code>	<code>evalml.pipelines.components.transformers.dimensionality_reduction.transformers</code>
module, 451	module, 645
<code>evalml.pipelines.components.estimators.column_selector_elastic</code>	<code>evalml.pipelines.components.transformers.encoding</code>
module, 454	module, 652
<code>evalml.pipelines.components.estimators.column_selector_elastic.compose</code>	<code>evalml.pipelines.components.transformers.encoding.compose</code>
module, 457	module, 652
<code>evalml.pipelines.components.estimators.column_selector_elastic.transformers</code>	<code>evalml.pipelines.components.transformers.encoding.transformers</code>
module, 460	module, 657
<code>evalml.pipelines.components.estimators.column_selector_lightgbm</code>	<code>evalml.pipelines.components.transformers.feature_selection</code>
module, 463	module, 666
<code>evalml.pipelines.components.estimators.column_selector_logistic_regression</code>	<code>evalml.pipelines.components.transformers.feature_selection.compose</code>
module, 467	module, 666
<code>evalml.pipelines.components.estimators.column_selector_logistic_regression.compose</code>	<code>evalml.pipelines.components.transformers.feature_selection.transformers</code>
module, 470	module, 668
<code>evalml.pipelines.components.estimators.column_selector_logistic_regression.transformers</code>	<code>evalml.pipelines.components.transformers.feature_selection.transformers</code>
module, 472	module, 672
<code>evalml.pipelines.components.estimators.column_selector_knn</code>	<code>evalml.pipelines.components.transformers.imputers</code>
module, 475	module, 683
<code>evalml.pipelines.components.estimators.column_selector_knn.compose</code>	<code>evalml.pipelines.components.transformers.imputers.compose</code>
module, 477	module, 683
<code>evalml.pipelines.components.estimators.column_selector_knn.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers</code>
module, 509	module, 686
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression</code>	<code>evalml.pipelines.components.transformers.imputers.transformers</code>
module, 509	module, 689
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.compose</code>
module, 513	module, 692
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers</code>
module, 515	module, 705
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.compose</code>
module, 518	module, 705
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers</code>
module, 522	module, 708
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.compose</code>
module, 524	module, 710
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers</code>
module, 528	module, 713
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers.compose</code>
module, 531	module, 715
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers.transformers</code>
module, 534	module, 718
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers.transformers.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers.transformers.compose</code>
module, 536	module, 721
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers.transformers.transformers</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers.transformers.transformers</code>
module, 539	module, 723
<code>evalml.pipelines.components.estimators.column_selector_parallel_regression.transformers.transformers.transformers.transformers.transformers.transformers.compose</code>	<code>evalml.pipelines.components.transformers.imputers.transformers.transformers.transformers.transformers.transformers.transformers.compose</code>
module, 542	module, 744

<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 744	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1180
<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 749	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1181
<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 757	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1182
<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 770	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1187
<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 770	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1187
<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 781	<code>evalml.pipelines.components.transformer_sampler_base_sampler</code>	module, 1187
<code>evalml.pipelines.components.utils</code>	<code>evalml.pipelines.components.utils</code>	module, 851	<code>evalml.pipelines.components.utils</code>	module, 1189
<code>evalml.pipelines.multiclass_classification_pipeline</code>	<code>evalml.pipelines.multiclass_classification_pipeline</code>	module, 1000	<code>evalml.pipelines.multiclass_classification_pipeline</code>	module, 1192
<code>evalml.pipelines.pipeline_base</code>	<code>evalml.pipelines.pipeline_base</code>	module, 1004	<code>evalml.pipelines.pipeline_base</code>	module, 1193
<code>evalml.pipelines.pipeline_meta</code>	<code>evalml.pipelines.pipeline_meta</code>	module, 1009	<code>evalml.pipelines.pipeline_meta</code>	module, 1193
<code>evalml.pipelines.regression_pipeline</code>	<code>evalml.pipelines.regression_pipeline</code>	module, 1010	<code>evalml.pipelines.regression_pipeline</code>	module, 1193
<code>evalml.pipelines.time_series_classification_pipeline</code>	<code>evalml.pipelines.time_series_classification_pipeline</code>	module, 1014	<code>evalml.pipelines.time_series_classification_pipeline</code>	module, 1193
<code>evalml.pipelines.time_series_regression_pipeline</code>	<code>evalml.pipelines.time_series_regression_pipeline</code>	module, 1028	<code>evalml.pipelines.time_series_regression_pipeline</code>	module, 1193
<code>evalml.pipelines.utils</code>	<code>evalml.pipelines.utils</code>	module, 1032	<code>evalml.pipelines.utils</code>	module, 1193
<code>evalml.preprocessing</code>	<code>evalml.preprocessing</code>	module, 1160	<code>evalml.preprocessing</code>	module, 1193
<code>evalml.preprocessing.data_splitters</code>	<code>evalml.preprocessing.data_splitters</code>	module, 1160	<code>evalml.preprocessing.data_splitters</code>	module, 1193
<code>evalml.preprocessing.data_splitters.balance_sampler</code>	<code>evalml.preprocessing.data_splitters.balance_sampler</code>	module, 1160	<code>evalml.preprocessing.data_splitters.balance_sampler</code>	module, 1193
<code>evalml.preprocessing.data_splitters.sample_sampler</code>	<code>evalml.preprocessing.data_splitters.sample_sampler</code>	module, 1161	<code>evalml.preprocessing.data_splitters.sample_sampler</code>	module, 1193
<code>evalml.preprocessing.data_splitters.time_series_split</code>	<code>evalml.preprocessing.data_splitters.time_series_split</code>	module, 1162	<code>evalml.preprocessing.data_splitters.time_series_split</code>	module, 1193
<code>evalml.preprocessing.data_splitters.training_validation_split</code>	<code>evalml.preprocessing.data_splitters.training_validation_split</code>	module, 1163	<code>evalml.preprocessing.data_splitters.training_validation_split</code>	module, 1193
<code>evalml.preprocessing.utils</code>	<code>evalml.preprocessing.utils</code>	module, 1167	<code>evalml.preprocessing.utils</code>	module, 1193
<code>evalml.problem_types</code>	<code>evalml.problem_types</code>	module, 1172	<code>evalml.problem_types</code>	module, 1193
<code>evalml.problem_types.problem_types</code>	<code>evalml.problem_types.problem_types</code>	module, 1172	<code>evalml.problem_types.problem_types</code>	module, 1193
<code>evalml.problem_types.utils</code>	<code>evalml.problem_types.utils</code>	module, 1173	<code>evalml.problem_types.utils</code>	module, 1193
<code>evalml.tuners</code>	<code>evalml.tuners</code>	module, 1177	<code>evalml.tuners</code>	module, 1193
<code>evalml.tuners.grid_search_tuner</code>	<code>evalml.tuners.grid_search_tuner</code>	module, 1177	<code>evalml.tuners.grid_search_tuner</code>	module, 1193
<code>evalml.tuners.random_search_tuner</code>	<code>evalml.tuners.random_search_tuner</code>	module, 1178	<code>evalml.tuners.random_search_tuner</code>	module, 1193

evalml.pipelines.components.estimators.classifiers),
 489 *feature_importance()*
 ExtraTreesClassifier (class in *evalml.pipelines.components*),
evalml.pipelines.components.estimators.classifiers.et_classifier
 457 *feature_importance()*
 ExtraTreesRegressor (class in *evalml.pipelines*),
 1074 *feature_importance()*
 ExtraTreesRegressor (class in *evalml.pipelines.components*), 899
 ExtraTreesRegressor (class in *evalml.pipelines.components.estimators*),
 609 *feature_importance()*
 ExtraTreesRegressor (class in *evalml.pipelines.components.estimators.regressors*),
 558 *feature_importance()*
 ExtraTreesRegressor (class in *evalml.pipelines.components.estimators.regressors*),
 525 *feature_importance()*
 F
 F1 (class in *evalml.objectives*), 371
 F1 (class in *evalml.objectives.standard_metrics*), 302
 F1Macro (class in *evalml.objectives*), 373
 F1Macro (class in *evalml.objectives.standard_metrics*),
 305
 F1Micro (class in *evalml.objectives*), 375
 F1Micro (class in *evalml.objectives.standard_metrics*),
 306
 F1Weighted (class in *evalml.objectives*), 376
 F1Weighted (class in *evalml.objectives.standard_metrics*), 308
 feature_importance()
evalml.pipelines.ARIMARegressor *property*),
 1036
 feature_importance()
evalml.pipelines.binary_classification_pipeline.BinaryClassifier
property), 988
 feature_importance()
evalml.pipelines.BinaryClassificationPipeline
property), 1039
 feature_importance()
evalml.pipelines.CatBoostClassifier *prop-*
erty), 1043
 feature_importance()
evalml.pipelines.CatBoostRegressor *prop-*
erty), 1046
 feature_importance()
evalml.pipelines.classification_pipeline.ClassificationPipeline
property), 994
 feature_importance()
evalml.pipelines.ClassificationPipeline *prop-*
erty), 1048
 feature_importance()
evalml.pipelines.components.ARIMARegressor

property), 859
feature_importance()
evalml.pipelines.components.BaselineClassifier
property), 861
feature_importance()
evalml.pipelines.components.BaselineRegressor
property), 864
feature_importance()
evalml.pipelines.components.CatBoostClassifier
property), 866
feature_importance()
evalml.pipelines.components.CatBoostRegressor
property), 869
feature_importance()
evalml.pipelines.components.DecisionTreeClassifier
property), 876
feature_importance()
evalml.pipelines.components.DecisionTreeRegressor
property), 879
feature_importance()
evalml.pipelines.components.ElasticNetClassifier
property), 891
feature_importance()
evalml.pipelines.components.ElasticNetRegressor
property), 893
feature_importance()
evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase
property), 429
feature_importance()
evalml.pipelines.components.ensemble.stacked_ensemble_classifier.StackedEnsembleClassifier
property), 432
feature_importance()
evalml.pipelines.components.ensemble.stacked_ensemble_regressor.StackedEnsembleRegressor
property), 435
feature_importance()
evalml.pipelines.components.ensemble.StackedEnsembleBase
property), 438
feature_importance()
evalml.pipelines.components.ensemble.StackedEnsembleClassifier
property), 441
feature_importance()
evalml.pipelines.components.ensemble.StackedEnsembleRegressor
property), 444
feature_importance()
evalml.pipelines.components.Estimator
property), 895
feature_importance()
evalml.pipelines.components.estimators.ARIMARegressor
property), 582
feature_importance()
evalml.pipelines.components.estimators.BaselineClassifier
property), 584
feature_importance()
evalml.pipelines.components.estimators.BaselineRegressor

`property)`, 587
`feature_importance()`
`(evalml.pipelines.components.estimators.CatBoostClassifier`
`property)`, 589
`feature_importance()`
`(evalml.pipelines.components.estimators.CatBoostRegressor`
`property)`, 592
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.baseline_classifier`
`property)`, 446
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.BaselineClassifier`
`property)`, 480
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.catboost_classifier`
`property)`, 450
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.CatBoostClassifier`
`property)`, 483
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.decision_tree_classifier`
`property)`, 453
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier`
`property)`, 486
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.elasticnet_classifier`
`property)`, 456
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`
`property)`, 489
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.et_classifier`
`property)`, 459
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier`
`property)`, 492
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.kneighbors_classifier`
`property)`, 462
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier`
`property)`, 495
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.lightgbm_classifier`
`property)`, 466
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.LightGBMClassifier`
`property)`, 498
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier`
`property)`, 469
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier`

`property)`, 501
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.RandomForestClassifier`
`property)`, 503
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.rf_classifier`
`property)`, 471
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.svm_classifier`
`property)`, 474
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.SVMClassifier`
`property)`, 506
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.xgboost_classifier`
`property)`, 477
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.XGBoostClassifier`
`property)`, 508
`feature_importance()`
`(evalml.pipelines.components.estimators.classifiers.XGBoostClassifier`
`property)`, 595
`feature_importance()`
`(evalml.pipelines.components.estimators.DecisionTreeRegressor`
`property)`, 598
`feature_importance()`
`(evalml.pipelines.components.estimators.ElasticNetClassifier`
`property)`, 601
`feature_importance()`
`(evalml.pipelines.components.estimators.ElasticNetRegressor`
`property)`, 603
`feature_importance()`
`(evalml.pipelines.components.estimators.Estimator`
`property)`, 605
`feature_importance()`
`(evalml.pipelines.components.estimators.estimator.Estimator`
`property)`, 578
`feature_importance()`
`(evalml.pipelines.components.estimators.ExtraTreesClassifier`
`property)`, 608
`feature_importance()`
`(evalml.pipelines.components.estimators.ExtraTreesRegressor`
`property)`, 611
`feature_importance()`
`(evalml.pipelines.components.estimators.KNeighborsClassifier`
`property)`, 614
`feature_importance()`
`(evalml.pipelines.components.estimators.LightGBMClassifier`
`property)`, 617
`feature_importance()`
`(evalml.pipelines.components.estimators.LightGBMRegressor`
`property)`, 620
`feature_importance()`
`(evalml.pipelines.components.estimators.LinearRegressor`

`property)`, 622
`feature_importance()`
`(evalml.pipelines.components.estimators.LogisticRegressionClassifier)`
`property)`, 625
`feature_importance()`
`(evalml.pipelines.components.estimators.RandomForestClassifier)`
`property)`, 627
`feature_importance()`
`(evalml.pipelines.components.estimators.RandomForestRegressor)`
`property)`, 630
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.arima_model.ARIMARegressor)`
`property)`, 512
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.ARIMARegressor)`
`property)`, 547
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.baseline_model.BaselineRegressor)`
`property)`, 514
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.BaselineRegressor)`
`property)`, 549
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.catboost_model.CatBoostRegressor)`
`property)`, 517
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.CatBoostRegressor)`
`property)`, 552
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.decision_tree_model.DecisionTreeRegressor)`
`property)`, 521
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor)`
`property)`, 555
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.elasticnet_model.ElasticNetRegressor)`
`property)`, 523
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.ElasticNetRegressor)`
`property)`, 557
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.et_regressor.ExtraTreesRegressor)`
`property)`, 527
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor)`
`property)`, 560
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.lightgbm_model.LightGBMRegressor)`
`property)`, 530
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.LightGBMRegressor)`
`property)`, 563
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.linear_model.LinearRegressor)`
`property)`, 533
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.LinearRegressor)`
`property)`, 566
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.RandomForestRegressor)`
`property)`, 568
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.rf_regressor.RandomForestRegressor)`
`property)`, 535
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.svm_regressor.SVMRegressor)`
`property)`, 538
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.SVMRegressor)`
`property)`, 571
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.time_series_baseline_model.TimeSeriesBaselineRegressor)`
`property)`, 541
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineRegressor)`
`property)`, 573
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.xgboost_model.XGBoostRegressor)`
`property)`, 544
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.XGBoostRegressor)`
`property)`, 576
`feature_importance()`
`(evalml.pipelines.components.estimators.regressors.SVMClassifier)`
`property)`, 632
`feature_importance()`
`(evalml.pipelines.components.estimators.SVMRegressor)`
`property)`, 634
`feature_importance()`
`(evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator)`
`property)`, 637
`feature_importance()`
`(evalml.pipelines.components.estimators.XGBoostClassifier)`
`property)`, 639
`feature_importance()`
`(evalml.pipelines.components.estimators.XGBoostRegressor)`
`property)`, 642
`feature_importance()`
`(evalml.pipelines.components.ExtraTreesClassifier)`
`property)`, 898
`feature_importance()`
`(evalml.pipelines.components.ExtraTreesRegressor)`
`property)`, 901
`feature_importance()`
`(evalml.pipelines.components.KNeighborsClassifier)`
`property)`, 909
`feature_importance()`
`(evalml.pipelines.components.LightGBMClassifier)`
`property)`, 909

[property](#)), 912
[feature_importance\(\)](#)
 ([evalml.pipelines.components.LightGBMRegressor](#)
 [property](#)), 915
[feature_importance\(\)](#)
 ([evalml.pipelines.components.LinearRegressor](#)
 [property](#)), 919
[feature_importance\(\)](#)
 ([evalml.pipelines.components.LogisticRegressionClassifier](#)
 [property](#)), 922
[feature_importance\(\)](#)
 ([evalml.pipelines.components.RandomForestClassifier](#)
 [property](#)), 936
[feature_importance\(\)](#)
 ([evalml.pipelines.components.RandomForestRegressor](#)
 [property](#)), 938
[feature_importance\(\)](#)
 ([evalml.pipelines.components.StackedEnsembleClassifier](#)
 [property](#)), 957
[feature_importance\(\)](#)
 ([evalml.pipelines.components.StackedEnsembleRegressor](#)
 [property](#)), 960
[feature_importance\(\)](#)
 ([evalml.pipelines.components.SVMClassifier](#)
 [property](#)), 964
[feature_importance\(\)](#)
 ([evalml.pipelines.components.SVMRegressor](#)
 [property](#)), 966
[feature_importance\(\)](#)
 ([evalml.pipelines.components.TimeSeriesBaselineEstimator](#)
 [property](#)), 975
[feature_importance\(\)](#)
 ([evalml.pipelines.components.XGBoostClassifier](#)
 [property](#)), 983
[feature_importance\(\)](#)
 ([evalml.pipelines.components.XGBoostRegressor](#)
 [property](#)), 985
[feature_importance\(\)](#)
 ([evalml.pipelines.DecisionTreeClassifier](#)
 [property](#)), 1056
[feature_importance\(\)](#)
 ([evalml.pipelines.DecisionTreeRegressor](#)
 [property](#)), 1059
[feature_importance\(\)](#)
 ([evalml.pipelines.ElasticNetClassifier](#)
 [property](#)), 1066
[feature_importance\(\)](#)
 ([evalml.pipelines.ElasticNetRegressor](#)
 [property](#)), 1068
[feature_importance\(\)](#)
 ([evalml.pipelines.Estimator](#) [property](#)), 1070
[feature_importance\(\)](#)
 ([evalml.pipelines.ExtraTreesClassifier](#)
 [property](#)), 1074
[feature_importance\(\)](#)
 ([evalml.pipelines.ExtraTreesRegressor](#)
 [property](#)), 1077
[feature_importance\(\)](#)
 ([evalml.pipelines.KNeighborsClassifier](#)
 [property](#)), 1082
[feature_importance\(\)](#)
 ([evalml.pipelines.LightGBMClassifier](#)
 [property](#)), 1084
[feature_importance\(\)](#)
 ([evalml.pipelines.LightGBMRegressor](#)
 [property](#)), 1087
[feature_importance\(\)](#)
 ([evalml.pipelines.LinearRegressor](#)
 [property](#)), 1089
[feature_importance\(\)](#)
 ([evalml.pipelines.LogisticRegressionClassifier](#)
 [property](#)), 1092
[feature_importance\(\)](#)
 ([evalml.pipelines.multiclass_classification_pipeline.MulticlassCl](#)
 [property](#)), 1002
[feature_importance\(\)](#)
 ([evalml.pipelines.MulticlassClassificationPipeline](#)
 [property](#)), 1095
[feature_importance\(\)](#)
 ([evalml.pipelines.pipeline_base.PipelineBase](#)
 [property](#)), 1006
[feature_importance\(\)](#)
 ([evalml.pipelines.PipelineBase](#)
 [property](#)), 1104
[feature_importance\(\)](#)
 ([evalml.pipelines.RandomForestClassifier](#)
 [property](#)), 1107
[feature_importance\(\)](#)
 ([evalml.pipelines.RandomForestRegressor](#)
 [property](#)), 1110
[feature_importance\(\)](#)
 ([evalml.pipelines.regression_pipeline.RegressionPipeline](#)
 [property](#)), 1012
[feature_importance\(\)](#)
 ([evalml.pipelines.RegressionPipeline](#)
 [property](#)), 1112
[feature_importance\(\)](#)
 ([evalml.pipelines.StackedEnsembleClassifier](#)
 [property](#)), 1124
[feature_importance\(\)](#)
 ([evalml.pipelines.StackedEnsembleRegressor](#)
 [property](#)), 1126
[feature_importance\(\)](#)
 ([evalml.pipelines.SVMClassifier](#)
 [property](#)), 1131
[feature_importance\(\)](#)
 ([evalml.pipelines.SVMRegressor](#)
 [property](#)), 1133

[feature_importance\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) property), 1017
[feature_importance\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline](#) property), 1021
[feature_importance\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline](#) property), 1026
[feature_importance\(\)](#) ([evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline](#) property), 1030
[feature_importance\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) property), 1138
[feature_importance\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#) property), 1143
[feature_importance\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#) property), 1147
[feature_importance\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#) property), 1151
[feature_importance\(\)](#) ([evalml.pipelines.XGBoostClassifier](#) property), 1157
[feature_importance\(\)](#) ([evalml.pipelines.XGBoostRegressor](#) property), 1159
[FeatureSelector](#) (class in [evalml.pipelines](#)), 1077
[FeatureSelector](#) (class in [evalml.pipelines.components](#)), 902
[FeatureSelector](#) (class in [evalml.pipelines.components.transformers](#)), 798
[FeatureSelector](#) (class in [evalml.pipelines.components.transformers.feature_selection](#)), 675
[FeatureSelector](#) (class in [evalml.pipelines.components.transformers.feature_selection.transformers](#)), 666
[fit\(\)](#) ([evalml.pipelines.ARIMAREgressor](#) method), 1036
[fit\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) method), 988
[fit\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) method), 1039
[fit\(\)](#) ([evalml.pipelines.CatBoostClassifier](#) method), 1043
[fit\(\)](#) ([evalml.pipelines.CatBoostRegressor](#) method), 1046
[fit\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) method), 994
[fit\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) method), 949
[fit\(\)](#) ([evalml.pipelines.component_graph.ComponentGraph](#) method), 998
[fit\(\)](#) ([evalml.pipelines.components.ARIMAREgressor](#) method), 850
[fit\(\)](#) ([evalml.pipelines.components.BaselineClassifier](#) method), 861
[fit\(\)](#) ([evalml.pipelines.components.BaselineRegressor](#) method), 864
[fit\(\)](#) ([evalml.pipelines.components.CatBoostClassifier](#) method), 866
[fit\(\)](#) ([evalml.pipelines.components.CatBoostRegressor](#) method), 869
[fit\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#) method), 849
[fit\(\)](#) ([evalml.pipelines.components.ComponentBase](#) method), 871
[fit\(\)](#) ([evalml.pipelines.components.DateTimeFeaturizer](#) method), 873
[fit\(\)](#) ([evalml.pipelines.components.DecisionTreeClassifier](#) method), 876
[fit\(\)](#) ([evalml.pipelines.components.DecisionTreeRegressor](#) method), 879
[fit\(\)](#) ([evalml.pipelines.components.DelayedFeatureTransformer](#) method), 881
[fit\(\)](#) ([evalml.pipelines.components.DFSTransformer](#) method), 884
[fit\(\)](#) ([evalml.pipelines.components.DropColumns](#) method), 886
[fit\(\)](#) ([evalml.pipelines.components.DropNullColumns](#) method), 888
[fit\(\)](#) ([evalml.pipelines.components.ElasticNetClassifier](#) method), 891
[fit\(\)](#) ([evalml.pipelines.components.ElasticNetRegressor](#) method), 893
[fit\(\)](#) ([evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase](#) method), 432
[fit\(\)](#) ([evalml.pipelines.components.ensemble.stacked_ensemble_classifier.StackedEnsembleClassifier](#) method), 441
[fit\(\)](#) ([evalml.pipelines.components.ensemble.stacked_ensemble_regressor.StackedEnsembleRegressor](#) method), 444
[fit\(\)](#) ([evalml.pipelines.components.Estimator](#) method), 895
[fit\(\)](#) ([evalml.pipelines.components.estimators.ARIMAREgressor](#) method), 582

`fit()` (`evalml.pipelines.components.estimators.BaselineClassifier` (`evalml.pipelines.components.estimators.DecisionTreeRegressor` method), 584 method), 598
`fit()` (`evalml.pipelines.components.estimators.BaselineRegressor` (`evalml.pipelines.components.estimators.ElasticNetClassifier` method), 587 method), 601
`fit()` (`evalml.pipelines.components.estimators.CatBoostClassifier` (`evalml.pipelines.components.estimators.ElasticNetRegressor` method), 589 method), 603
`fit()` (`evalml.pipelines.components.estimators.CatBoostRegressor` (`evalml.pipelines.components.estimators.Estimator` method), 592 method), 605
`fit()` (`evalml.pipelines.components.estimators.classifiers.BaselineClassifier` (`evalml.pipelines.components.estimators.estimator.Estimator` method), 446 method), 578
`fit()` (`evalml.pipelines.components.estimators.classifiers.BaselineClassifier` (`evalml.pipelines.components.estimators.ExtraTreesClassifier` method), 480 method), 608
`fit()` (`evalml.pipelines.components.estimators.classifiers.CatBoostClassifier` (`evalml.pipelines.components.estimators.ExtraTreesRegressor` method), 450 method), 611
`fit()` (`evalml.pipelines.components.estimators.classifiers.CatBoostClassifier` (`evalml.pipelines.components.estimators.KNeighborsClassifier` method), 483 method), 614
`fit()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` (`evalml.pipelines.components.estimators.LightGBMClassifier` method), 453 method), 617
`fit()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` (`evalml.pipelines.components.estimators.LightGBMRegressor` method), 486 method), 620
`fit()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier` (`evalml.pipelines.components.estimators.LinearRegressor` method), 456 method), 622
`fit()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier` (`evalml.pipelines.components.estimators.LogisticRegressionClassifier` method), 489 method), 625
`fit()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier` (`evalml.pipelines.components.estimators.RandomForestClassifier` method), 459 method), 627
`fit()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier` (`evalml.pipelines.components.estimators.RandomForestRegressor` method), 492 method), 630
`fit()` (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier` (`evalml.pipelines.components.estimators.regressors.arima_regressor` method), 462 method), 512
`fit()` (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier` (`evalml.pipelines.components.estimators.regressors.ARIMARegressor` method), 495 method), 547
`fit()` (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier` (`evalml.pipelines.components.estimators.regressors.baseline_regressor` method), 466 method), 514
`fit()` (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier` (`evalml.pipelines.components.estimators.regressors.BaselineRegressor` method), 498 method), 549
`fit()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` (`evalml.pipelines.components.estimators.regressors.catboost_regressor` method), 469 method), 517
`fit()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier` (`evalml.pipelines.components.estimators.regressors.CatBoostRegressor` method), 501 method), 552
`fit()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` (`evalml.pipelines.components.estimators.regressors.decision_tree_regressor` method), 503 method), 521
`fit()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` (`evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor` method), 471 method), 555
`fit()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` (`evalml.pipelines.components.estimators.regressors.elasticnet_regressor` method), 474 method), 523
`fit()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` (`evalml.pipelines.components.estimators.regressors.ElasticNetRegressor` method), 506 method), 557
`fit()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` (`evalml.pipelines.components.estimators.regressors.et_regressor.ExtraTreesRegressor` method), 477 method), 527
`fit()` (`evalml.pipelines.components.estimators.classifiers.XGBoostClassifier` (`evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor` method), 508 method), 560
`fit()` (`evalml.pipelines.components.estimators.DecisionTreeClassifier` (`evalml.pipelines.components.estimators.regressors.lightgbm_regressor` method), 595 method), 530

[fit \(\) \(evalml.pipelines.components.estimators.regressors.LightGBMRegressor method\), 563](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.LinearRegressor method\), 533](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.LikeARegression method\), 566](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.RandomForestRegressor method\), 568](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.RidgeRegressor method\), 535](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.SVMRegressor method\), 538](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.SVMRegressor method\), 571](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator method\), 541](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineEstimator method\), 573](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.XGBoostRegressor method\), 544](#)
[fit \(\) \(evalml.pipelines.components.estimators.regressors.XGBoostRegressor method\), 576](#)
[fit \(\) \(evalml.pipelines.components.estimators.SVMClassifier method\), 632](#)
[fit \(\) \(evalml.pipelines.components.estimators.SVMRegressor method\), 634](#)
[fit \(\) \(evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator method\), 637](#)
[fit \(\) \(evalml.pipelines.components.estimators.XGBoostClassifier method\), 639](#)
[fit \(\) \(evalml.pipelines.components.estimators.XGBoostRegressor method\), 642](#)
[fit \(\) \(evalml.pipelines.components.ExtraTreesClassifier method\), 898](#)
[fit \(\) \(evalml.pipelines.components.ExtraTreesRegressor method\), 901](#)
[fit \(\) \(evalml.pipelines.components.FeatureSelector method\), 903](#)
[fit \(\) \(evalml.pipelines.components.Imputer method\), 906](#)
[fit \(\) \(evalml.pipelines.components.KNeighborsClassifier method\), 909](#)
[fit \(\) \(evalml.pipelines.components.LightGBMClassifier method\), 912](#)
[fit \(\) \(evalml.pipelines.components.LightGBMRegressor method\), 915](#)
[fit \(\) \(evalml.pipelines.components.LinearDiscriminantAnalysis method\), 917](#)
[fit \(\) \(evalml.pipelines.components.LinearRegressor method\), 919](#)
[fit \(\) \(evalml.pipelines.components.LogisticRegressionClassifier method\), 922](#)
[fit \(\) \(evalml.pipelines.components.LSA method\), 924](#)
[fit \(\) \(evalml.pipelines.components.OneHotEncoder method\), 926](#)
[fit \(\) \(evalml.pipelines.components.PCA method\), 929](#)
[fit \(\) \(evalml.pipelines.components.PerColumnImputer method\), 931](#)
[fit \(\) \(evalml.pipelines.components.PolynomialDetrender method\), 933](#)
[fit \(\) \(evalml.pipelines.components.RandomForestClassifier method\), 936](#)
[fit \(\) \(evalml.pipelines.components.RandomForestRegressor method\), 938](#)
[fit \(\) \(evalml.pipelines.components.RFClassifierSelectFromModel method\), 941](#)
[fit \(\) \(evalml.pipelines.components.RFRegressorSelectFromModel method\), 943](#)
[fit \(\) \(evalml.pipelines.components.SelectFromModel method\), 946](#)
[fit \(\) \(evalml.pipelines.components.SimpleImputer method\), 948](#)
[fit \(\) \(evalml.pipelines.components.SMOTEClassifier method\), 950](#)
[fit \(\) \(evalml.pipelines.components.SMOTENCSampler method\), 952](#)
[fit \(\) \(evalml.pipelines.components.SMOTESampler method\), 955](#)
[fit \(\) \(evalml.pipelines.components.StackedEnsembleClassifier method\), 957](#)
[fit \(\) \(evalml.pipelines.components.StackedEnsembleRegressor method\), 960](#)
[fit \(\) \(evalml.pipelines.components.StandardScaler method\), 962](#)
[fit \(\) \(evalml.pipelines.components.SVMClassifier method\), 964](#)
[fit \(\) \(evalml.pipelines.components.SVMRegressor method\), 966](#)
[fit \(\) \(evalml.pipelines.components.TargetEncoder method\), 969](#)
[fit \(\) \(evalml.pipelines.components.TargetImputer method\), 971](#)
[fit \(\) \(evalml.pipelines.components.TextFeaturizer method\), 973](#)
[fit \(\) \(evalml.pipelines.components.TimeSeriesBaselineEstimator method\), 975](#)
[fit \(\) \(evalml.pipelines.components.Transformer method\), 977](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selectors.ColumnSelector method\), 776](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selectors.DropColumns method\), 778](#)
[fit \(\) \(evalml.pipelines.components.transformers.column_selectors.SelectColumns method\), 780](#)
[fit \(\) \(evalml.pipelines.components.transformers.DateTimeFeaturizer method\), 788](#)
[fit \(\) \(evalml.pipelines.components.transformers.DelayedFeatureTransformer method\), 791](#)

`fit()` (`evalml.pipelines.components.transformers.DFSTransformer` method), 793
`fit()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis` method), 804
`fit()` (`evalml.pipelines.components.transformers.dimensionality_reduction.LinearDiscriminantAnalysis` method), 806
`fit()` (`evalml.pipelines.components.transformers.dimensionality_reduction.LinearDiscriminantAnalysis` method), 809
`fit()` (`evalml.pipelines.components.transformers.dimensionality_reduction.PCA` method), 811
`fit()` (`evalml.pipelines.components.transformers.dimensionality_reduction.PCA` method), 813
`fit()` (`evalml.pipelines.components.transformers.DropColumns` method), 795
`fit()` (`evalml.pipelines.components.transformers.DropNullColumns` method), 797
`fit()` (`evalml.pipelines.components.transformers.encoders.OneHotEncoder` method), 655
`fit()` (`evalml.pipelines.components.transformers.encoders.OneHotEncoder` method), 727
`fit()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder` method), 662
`fit()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder` method), 709
`fit()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder` method), 730
`fit()` (`evalml.pipelines.components.transformers.encoders.TargetEncoder` method), 732
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 667
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 712
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 734
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 714
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 736
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 717
`fit()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector` method), 719
`fit()` (`evalml.pipelines.components.transformers.FeatureSelector` method), 799
`fit()` (`evalml.pipelines.components.transformers.FeatureSelector` method), 739
`fit()` (`evalml.pipelines.components.transformers.Imputer` method), 802
`fit()` (`evalml.pipelines.components.transformers.imputers.Imputer` method), 696
`fit()` (`evalml.pipelines.components.transformers.imputers.Imputer` method), 724
`fit()` (`evalml.pipelines.components.transformers.imputers.Imputer` method), 741
`fit()` (`evalml.pipelines.components.transformers.imputers.Imputer` method), 743
`fit()` (`evalml.pipelines.components.transformers.imputers.PerColumnImputer` method), 699
`fit()` (`evalml.pipelines.components.transformers.imputers.PerColumnImputer` method), 818
`fit()` (`evalml.pipelines.components.transformers.imputers.SimpleImputer` method), 690
`fit()` (`evalml.pipelines.components.transformers.imputers.SimpleImputer` method), 821
`fit()` (`evalml.pipelines.components.transformers.imputers.SimpleImputer` method), 746
`fit()` (`evalml.pipelines.components.transformers.imputers.TargetImputer` method), 693
`fit()` (`evalml.pipelines.components.transformers.imputers.TargetImputer` method), 748
`fit()` (`evalml.pipelines.components.transformers.imputers.TargetImputer` method), 751

[fit\(\) \(evalml.pipelines.StandardScaler method\), 1128](#)
[fit\(\) \(evalml.pipelines.SVMClassifier method\), 1131](#)
[fit\(\) \(evalml.pipelines.SVMRegressor method\), 1133](#)
[fit\(\) \(evalml.pipelines.TargetEncoder method\), 1135](#)
[fit\(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline method\), 1017](#)
[fit\(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method\), 1021](#)
[fit\(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline method\), 1026](#)
[fit\(\) \(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method\), 1030](#)
[fit\(\) \(evalml.pipelines.TimeSeriesBinaryClassificationPipeline method\), 1138](#)
[fit\(\) \(evalml.pipelines.TimeSeriesClassificationPipeline method\), 1143](#)
[fit\(\) \(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method\), 1147](#)
[fit\(\) \(evalml.pipelines.TimeSeriesRegressionPipeline method\), 1151](#)
[fit\(\) \(evalml.pipelines.Transformer method\), 1154](#)
[fit\(\) \(evalml.pipelines.XGBoostClassifier method\), 1157](#)
[fit\(\) \(evalml.pipelines.XGBoostRegressor method\), 1159](#)
[fit_features\(\) \(evalml.pipelines.component_graph.ComponentGraph method\), 998](#)
[fit_features\(\) \(evalml.pipelines.ComponentGraph method\), 1052](#)
[fit_resample\(\) \(evalml.preprocessing.data_splitters.balanced_classification_sampler.BalancedClassificationSampler method\), 1161](#)
[fit_resample\(\) \(evalml.preprocessing.data_splitters.BalancedClassificationSampler method\), 1165](#)
[fit_resample\(\) \(evalml.preprocessing.data_splitters.sampler_base.SamplerBase method\), 1161](#)
[fit_resample\(\) \(evalml.preprocessing.data_splitters.SamplerBase method\), 1165](#)
[fit_transform\(\) \(evalml.pipelines.components.DateTimeFeaturizer method\), 873](#)
[fit_transform\(\) \(evalml.pipelines.components.DelayedFeatureTransformer method\), 882](#)
[fit_transform\(\) \(evalml.pipelines.components.DFSTransformer method\), 884](#)
[fit_transform\(\) \(evalml.pipelines.components.DropColumns method\), 886](#)
[fit_transform\(\) \(evalml.pipelines.components.DropNullColumns method\), 888](#)
[fit_transform\(\) \(evalml.pipelines.components.FeatureSelector method\), 903](#)
[fit_transform\(\) \(evalml.pipelines.components.Imputer method\), 906](#)
[fit_transform\(\) \(evalml.pipelines.components.LinearDiscriminantAnalysis method\), 917](#)
[fit_transform\(\) \(evalml.pipelines.components.LSA method\), 924](#)
[fit_transform\(\) \(evalml.pipelines.components.OneHotEncoder method\), 927](#)
[fit_transform\(\) \(evalml.pipelines.components.PCA method\), 929](#)
[fit_transform\(\) \(evalml.pipelines.components.PerColumnImputer method\), 934](#)
[fit_transform\(\) \(evalml.pipelines.components.PolynomialDetrender method\), 938](#)
[fit_transform\(\) \(evalml.pipelines.components.RFClassifierSelectFromModel method\), 943](#)
[fit_transform\(\) \(evalml.pipelines.components.RFRegressorSelectFromModel method\), 944](#)
[fit_transform\(\) \(evalml.pipelines.components.SelectColumns method\), 946](#)
[fit_transform\(\) \(evalml.pipelines.components.SimpleImputer method\), 948](#)
[fit_transform\(\) \(evalml.pipelines.components.SMOTENCSampler method\), 950](#)
[fit_transform\(\) \(evalml.pipelines.components.SMOTENSampler method\), 953](#)
[fit_transform\(\) \(evalml.pipelines.components.SMOTESampler method\), 955](#)
[fit_transform\(\) \(evalml.pipelines.components.StandardScaler method\), 962](#)
[fit_transform\(\) \(evalml.pipelines.components.TargetEncoder method\), 969](#)
[fit_transform\(\) \(evalml.pipelines.components.TargetImputer method\), 971](#)
[fit_transform\(\) \(evalml.pipelines.components.TargetNormalizer method\), 973](#)
[fit_transform\(\) \(evalml.pipelines.components.Transformer method\), 978](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.column_transformer.ColumnTransformer method\), 776](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.column_transformer.ColumnTransformer method\), 778](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.column_transformer.ColumnTransformer method\), 780](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.DateTimeFeaturizer method\), 788](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.DelayedFeatureTransformer method\), 791](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.DFSTransformer method\), 793](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.DimensionalityReduction method\), 644](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.DimensionalityReduction method\), 649](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.DimensionalityReduction method\), 651](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.DimensionalityReduction method\), 647](#)
[fit_transform\(\) \(evalml.pipelines.components.transformers.DropColumns method\), 647](#)

[get_component\(\) \(evalml.pipelines.regression_pipeline.RegressionPipeline method\), 1012](#)
[get_component\(\) \(evalml.pipelines.RegressionPipeline method\), 1113](#)
[get_component\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesBinaryClassificationPipeline method\), 1017](#)
[get_component\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method\), 1021](#)
[get_component\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline method\), 1026](#)
[get_component\(\) \(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method\), 1030](#)
[get_component\(\) \(evalml.pipelines.TimeSeriesBinaryClassificationPipeline method\), 1138](#)
[get_component\(\) \(evalml.pipelines.TimeSeriesClassificationPipeline method\), 1143](#)
[get_component\(\) \(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method\), 1147](#)
[get_component\(\) \(evalml.pipelines.TimeSeriesRegressionPipeline method\), 1151](#)
[get_core_objective_names\(\) \(in module evalml.objectives\), 381](#)
[get_core_objective_names\(\) \(in module evalml.objectives.utils\), 348](#)
[get_core_objectives\(\) \(in module evalml.objectives\), 381](#)
[get_core_objectives\(\) \(in module evalml.objectives.utils\), 348](#)
[get_default_primary_search_objective\(\) \(in module evalml.automl\), 181](#)
[get_default_primary_search_objective\(\) \(in module evalml.automl.utils\), 174](#)
[get_estimators\(\) \(evalml.pipelines.component_graph.ComponentGraph method\), 999](#)
[get_estimators\(\) \(evalml.pipelines.ComponentGraph method\), 1053](#)
[get_estimators\(\) \(in module evalml.pipelines.components.utils\), 852](#)
[get_evalml_root\(\) \(in module evalml.utils.cli_utils\), 1188](#)
[get_feature_names\(\) \(evalml.pipelines.components.DateTimeFeaturizer method\), 874](#)
[get_feature_names\(\) \(evalml.pipelines.components.OneHotEncoder method\), 927](#)
[get_feature_names\(\) \(evalml.pipelines.components.TargetEncoder method\), 969](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.DateTimeFeaturizer method\), 788](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder method\), 1012](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.encoders.OneHotEncoder method\), 662](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder method\), 665](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.preprocessing.datetime_featurizer.DateTimeFeaturizer method\), 707](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.preprocessing.DateTimeFeaturizer method\), 728](#)
[get_feature_names\(\) \(evalml.pipelines.components.transformers.TargetEncoder method\), 837](#)
[get_feature_names\(\) \(evalml.pipelines.OneHotEncoder method\), 1099](#)
[get_feature_names\(\) \(evalml.pipelines.TargetEncoder method\), 1135](#)
[get_hyperparameter_ranges\(\) \(in module evalml.automl.utils\), 174](#)
[get_importable_subclasses\(\) \(in module evalml.utils\), 1195](#)
[get_importable_subclasses\(\) \(in module evalml.utils.gen_utils\), 1190](#)
[get_installed_packages\(\) \(in module evalml.utils.cli_utils\), 1188](#)
[get_last_component\(\) \(evalml.pipelines.component_graph.ComponentGraph method\), 999](#)
[get_last_component\(\) \(evalml.pipelines.ComponentGraph method\), 1053](#)
[get_linear_coefficients\(\) \(in module evalml.model_understanding\), 258](#)
[get_linear_coefficients\(\) \(in module evalml.model_understanding.graphs\), 244](#)
[get_logger\(\) \(in module evalml.utils\), 1196](#)
[get_logger\(\) \(in module evalml.utils.logger\), 1193](#)
[get_n_splits\(\) \(evalml.preprocessing.data_splitters.time_series_split.TimeSeriesSplit method\), 1162](#)
[get_n_splits\(\) \(evalml.preprocessing.data_splitters.TimeSeriesSplit method\), 1166](#)
[get_n_splits\(\) \(evalml.preprocessing.data_splitters.training_validation_split.TrainingValidationSplit method\), 1015](#)

[1113](#) [\(evalml.pipelines.RegressionPipeline method\)](#),
[graph \(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline method\)](#), [1017](#) [graph_feature_importance \(\)](#)
[graph \(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method\)](#), [1022](#) [\(evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method\)](#), [1017](#)
[graph \(\) \(evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline method\)](#), [1026](#) [\(evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline method\)](#), [1026](#)
[graph \(\) \(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method\)](#), [1030](#) [graph_feature_importance \(\)](#)
[graph \(\) \(evalml.pipelines.TimeSeriesBinaryClassificationPipeline method\)](#), [1138](#) [\(evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline method\)](#), [1026](#)
[graph \(\) \(evalml.pipelines.TimeSeriesClassificationPipeline method\)](#), [1143](#) [graph_feature_importance \(\)](#)
[graph \(\) \(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method\)](#), [1030](#) [\(evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method\)](#), [1030](#)
[graph \(\) \(evalml.pipelines.TimeSeriesRegressionPipeline method\)](#), [1151](#) [\(evalml.pipelines.TimeSeriesBinaryClassificationPipeline method\)](#), [1139](#)
[graph_binary_objective_vs_threshold \(\)](#) [\(in module evalml.model_understanding\)](#), [258](#) [graph_feature_importance \(\)](#)
[graph_binary_objective_vs_threshold \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [245](#) [\(evalml.pipelines.TimeSeriesClassificationPipeline method\)](#), [1143](#)
[graph_confusion_matrix \(\)](#) [\(in module evalml.model_understanding\)](#), [259](#) [\(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method\)](#), [1147](#)
[graph_confusion_matrix \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [245](#) [\(evalml.pipelines.TimeSeriesRegressionPipeline method\)](#), [1151](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline method\)](#), [989](#) [graph_force_plot \(\)](#) [\(in module evalml.model_understanding.force_plots\)](#), [241](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.BinaryClassificationPipeline method\)](#), [1040](#) [graph_partial_dependence \(\)](#) [\(in module evalml.model_understanding\)](#), [259](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.classification_pipeline.ClassificationPipeline method\)](#), [994](#) [graph_partial_dependence \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [245](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.ClassificationPipeline method\)](#), [1049](#) [graph_permutation_importance \(\)](#) [\(in module evalml.model_understanding\)](#), [260](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method\)](#), [1003](#) [graph_permutation_importance \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [246](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.MulticlassClassificationPipeline method\)](#), [1095](#) [graph_precision_recall_curve \(\)](#) [\(in module evalml.model_understanding\)](#), [260](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.pipeline_base.PipelineBase method\)](#), [1007](#) [graph_prediction_vs_actual \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [247](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.PipelineBase method\)](#), [1104](#) [graph_prediction_vs_actual_over_time \(\)](#) [\(in module evalml.model_understanding\)](#), [261](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.regression_pipeline.RegressionPipeline method\)](#), [1013](#) [graph_prediction_vs_actual_over_time \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [247](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.regression_pipeline.RegressionPipeline method\)](#), [1013](#) [graph_roc_curve \(\)](#) [\(in module evalml.model_understanding\)](#), [261](#)
[graph_feature_importance \(\)](#) [\(evalml.pipelines.regression_pipeline.RegressionPipeline method\)](#), [1013](#) [graph_roc_curve \(\)](#) [\(in module evalml.model_understanding.graphs\)](#), [247](#)

[graph_t_sne\(\)](#) (in module [evalml.model_understanding](#)), 261
[graph_t_sne\(\)](#) (in module [evalml.model_understanding.graphs](#)), 248
[greater_is_better\(\)](#) ([evalml.objectives.binary_classification_objective.BinaryClassificationObjective](#) property), 266
[greater_is_better\(\)](#) ([evalml.objectives.BinaryClassificationObjective](#) property), 366
[greater_is_better\(\)](#) ([evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective](#) property), 277
[greater_is_better\(\)](#) ([evalml.objectives.MulticlassClassificationObjective](#) property), 401
[greater_is_better\(\)](#) ([evalml.objectives.objective_base.ObjectiveBase](#) property), 279
[greater_is_better\(\)](#) ([evalml.objectives.ObjectiveBase](#) property), 403
[greater_is_better\(\)](#) ([evalml.objectives.regression_objective.RegressionObjective](#) property), 281
[greater_is_better\(\)](#) ([evalml.objectives.RegressionObjective](#) property), 421
[greater_is_better\(\)](#) ([evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective](#) property), 346
[GridSearchTuner](#) (class in [evalml.tuners](#)), 1183
[GridSearchTuner](#) (class in [evalml.tuners.grid_search_tuner](#)), 1177

H

[handle_component_class\(\)](#) (in module [evalml.pipelines.components.utils](#)), 852
[handle_model_family\(\)](#) (in module [evalml.model_family](#)), 233
[handle_model_family\(\)](#) (in module [evalml.model_family.utils](#)), 233
[handle_problem_types\(\)](#) (in module [evalml.problem_types](#)), 1175
[handle_problem_types\(\)](#) (in module [evalml.problem_types.utils](#)), 1173
[HighlyNullDataCheck](#) (class in [evalml.data_checks](#)), 217
[HighlyNullDataCheck](#) (class in [evalml.data_checks.highly_null_data_check](#)), 195

I

[IDColumnsDataCheck](#) (class in [evalml.data_checks](#)),
 [IDColumnsDataCheck](#) (class in [evalml.data_checks.id_columns_data_check](#)), 196
[import_or_raise\(\)](#) (in module [evalml.utils](#)), 1196
[Imputer](#) (class in [evalml.pipelines.components](#)), 904
[Imputer](#) (class in [evalml.pipelines.components.transformers](#)), 800
[Imputer](#) (class in [evalml.pipelines.components.transformers.imputers](#)), 905
[Imputer](#) (class in [evalml.pipelines.components.transformers.imputers.imputer](#)), 684
[infer_feature_types\(\)](#) (in module [evalml.utils](#)), 1196
[infer_feature_types\(\)](#) (in module [evalml.utils.woodwork_utils](#)), 1193
[info\(\)](#) ([evalml.automl.engine.engine_base.JobLogger](#) method), 157
[instantiate\(\)](#) ([evalml.pipelines.component_graph.ComponentGraph](#) method), 999
[instantiate\(\)](#) ([evalml.pipelines.ComponentGraph](#) method), 1053
[InvalidTargetDataCheck](#) (class in [evalml.data_checks](#)), 218
[InvalidTargetDataCheck](#) (class in [evalml.data_checks.invalid_targets_data_check](#)), 197
[inverse_transform\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) method), 989
[inverse_transform\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) method), 1040
[inverse_transform\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) method), 995
[inverse_transform\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) method), 1049
[inverse_transform\(\)](#) ([evalml.pipelines.component_graph.ComponentGraph](#) method), 999
[inverse_transform\(\)](#) ([evalml.pipelines.ComponentGraph](#) method), 1053
[inverse_transform\(\)](#) ([evalml.pipelines.components.PolynomialDetrender](#) method), 934
[inverse_transform\(\)](#) ([evalml.pipelines.components.transformers.PolynomialDetrender](#) method), 816
[inverse_transform\(\)](#)

(evalml.pipelines.components.transformers.preprocessing.high_dimensionality_utils),
 method), 719

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.components.transformers.preprocessing.PolyomialObjectiveBinaryClassificationObjective.BinaryClassificationObjective), 739

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.components.transformers.transformer.TargetTransformerObjective.BinaryClassificationObjective), 782

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline), 1003

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.MulticlassClassificationPipeline (evalml.objectives.MulticlassClassificationObjective), 1095

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.pipeline_base.PipelineBase (evalml.objectives.objective_base.ObjectiveBase), 1007

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.PipelineBase method), (evalml.objectives.ObjectiveBase property), 1104

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.regression_pipeline.RegressionPipeline (evalml.objectives.regression_objective.RegressionObjective), 1013

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.RegressionPipeline method), (evalml.objectives.RegressionObjective property), 1113

inverse_transform() is_bounded_like_percentage()
 (evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline (evalml.objectives.TimeSeriesRegressionObjective), 1017

inverse_transform() is_cancelled() (evalml.automl.engine.cf_engine.CFComputation), 1022

inverse_transform() is_cancelled() (evalml.automl.engine.dask_engine.DaskComputation), 1026

inverse_transform() is_classification() (in module evalml.problem_types), 1175

inverse_transform() is_classification() (in module evalml.problem_types), 1174

inverse_transform() is_defined_for_problem_type()
 (evalml.pipelines.TimeSeriesBinaryClassificationPipeline method), 351

inverse_transform() is_defined_for_problem_type()
 (evalml.pipelines.TimeSeriesClassificationPipeline method), 353

inverse_transform() is_defined_for_problem_type()
 (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method), 1147

inverse_transform() is_defined_for_problem_type()
 (evalml.pipelines.TimeSeriesRegressionPipeline method), 1151

is_all_numeric() (in module evalml.utils), 1197

is_all_numeric() (in module evalml.utils.gen_utils), 1191

is_binary() (in module evalml.problem_types), 1175

is_bounded_like_percentage()
 (evalml.objectives.BinaryClassificationObjective), 266

is_bounded_like_percentage()
 (evalml.objectives.BinaryClassificationObjective), 366

is_bounded_like_percentage()
 (evalml.objectives.MulticlassClassificationObjective), 277

is_bounded_like_percentage()
 (evalml.objectives.MulticlassClassificationObjective), 401

is_bounded_like_percentage()
 (evalml.objectives.objective_base.ObjectiveBase), 279

is_bounded_like_percentage()
 (evalml.objectives.ObjectiveBase property), 403

is_bounded_like_percentage()
 (evalml.objectives.regression_objective.RegressionObjective), 282

is_bounded_like_percentage()
 (evalml.objectives.RegressionObjective property), 421

is_bounded_like_percentage()
 (evalml.objectives.TimeSeriesRegressionObjective), 346

is_cancelled() (evalml.automl.engine.cf_engine.CFComputation), 422

is_cancelled() (evalml.automl.engine.dask_engine.DaskComputation), 154

is_classification() (in module evalml.problem_types), 1175

is_classification() (in module evalml.problem_types), 1174

is_defined_for_problem_type()
 (evalml.objectives.AccuracyBinary class), 351

is_defined_for_problem_type()
 (evalml.objectives.AccuracyMulticlass class), 353

is_defined_for_problem_type()
 (evalml.objectives.AUC class method), 355

is_defined_for_problem_type()
 (evalml.objectives.AUCMacro class method), 357

is_defined_for_problem_type()
 (evalml.objectives.AUCMicro class method), 358

is_defined_for_problem_type()
 (evalml.objectives.AUCWeighted class method), 360

<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.BalancedAccuracyBinary</code> class method), 362	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MaxError</code> class method), 391
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.BalancedAccuracyMulticlass</code> class method), 364	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MCCBinary</code> class method), 393
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.binary_classification_objective.BinaryClassificationObjective</code> class method), 266	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MCCMulticlass</code> class method), 395
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.BinaryClassificationObjective</code> class method), 366	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MeanSquaredLogError</code> class method), 396
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.cost_benefit_matrix.CostBenefitMatrix</code> class method), 269	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MedianAE</code> class method), 398
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.CostBenefitMatrix</code> class method), 368	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MSE</code> class method), 400
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.ExpVariance</code> class method), 370	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.multiclass_classification_objective.MulticlassC</code> class method), 277
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.F1</code> class method), 372	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MulticlassClassificationObjective</code> class method), 401
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.F1Macro</code> class method), 374	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.objective_base.ObjectiveBase</code> class method), 279
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.F1Micro</code> class method), 376	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.ObjectiveBase</code> class method), 403
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.F1Weighted</code> class method), 377	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.Precision</code> class method), 406
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.fraud_cost.FraudCost</code> class method), 272	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionMacro</code> class method), 407
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.FraudCost</code> class method), 379	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionMicro</code> class method), 409
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.lead_scoring.LeadScoring</code> class method), 275	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.PrecisionWeighted</code> class method), 410
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LeadScoring</code> class method), 382	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.R2</code> class method), 412
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LogLossBinary</code> class method), 385	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.Recall</code> class method), 414
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.LogLossMulticlass</code> class method), 386	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RecallMacro</code> class method), 416
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MAE</code> class method), 388	<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.RecallMicro</code> class method), 417
<code>is_defined_for_problem_type()</code> (<code>evalml.objectives.MAPE</code> class method),	<code>is_defined_for_problem_type()</code>

<i>(evalml.objectives.RecallWeighted method), 419</i>	<i>class</i>	<i>(evalml.objectives.standard_metrics.F1Micro class method), 307</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.regression_objective.RegressionObjective class method), 282</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1Weighted class method), 308</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.RegressionObjective class method), 421</i>	<i>class</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.LogLossBinary class method), 310</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.RootMeanSquaredError class method), 422</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.LogLossMulticlass class method), 312</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.RootMeanSquaredLogError class method), 424</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MAE class method), 314</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.sensitivity_low_alert.SensitivityLowAlert class method), 284</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MAPE class method), 315</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.SensitivityLowAlert class method), 426</i>	<i>class</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MaxError class method), 317</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AccuracyBinary class method), 287</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MCCBinary class method), 319</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AccuracyMulticlass class method), 289</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MCCMulticlass class method), 321</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUC class method), 291</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MeanSquaredLogError class method), 322</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCMacro class method), 293</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MedianAE class method), 324</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCMicro class method), 295</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.MSE class method), 325</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.AUCWeighted class method), 296</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.Precision class method), 327</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.BalancedAccuracyBinary class method), 298</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionMacro class method), 329</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.BalancedAccuracyMulticlass class method), 300</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionMicro class method), 331</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.ExpVariance class method), 302</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.PrecisionWeighted class method), 332</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1 class method), 304</i>	<i>class</i>	<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.R2 class method), 334</i>
<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.F1Macro class method), 305</i>		<i>is_defined_for_problem_type()</i> <i>(evalml.objectives.standard_metrics.Recall class method), 336</i>
<i>is_defined_for_problem_type()</i>		<i>is_defined_for_problem_type()</i>

(*evalml.objectives.standard_metrics.RecallMacro* class method), 338

(*evalml.model_family.ModelFamily* method), 234

is_defined_for_problem_type() (*evalml.objectives.standard_metrics.RecallMicro* class method), 339

IterativeAlgorithm (class in *evalml.automl.automl_algorithm*), 149

is_defined_for_problem_type() (*evalml.objectives.standard_metrics.RecallWeighted* class method), 341

IterativeAlgorithm (class in *evalml.automl.automl_algorithm.iterative_algorithm*), 146

J

is_defined_for_problem_type() (*evalml.objectives.standard_metrics.RootMeanSquaredError* class method), 342

RootMeanSquaredError (class in *evalml.automl.engine.engine_base*), 157

is_defined_for_problem_type() (*evalml.objectives.standard_metrics.RootMeanSquaredLogError* class method), 344

jupyter_check() (in module *evalml.utils*), 1197

RootMeanSquaredLogError (class in *evalml.automl.engine.engine_base*), 157

jupyter_check() (in module *evalml.utils.gen_utils*), 1191

K

is_defined_for_problem_type() (*evalml.objectives.time_series_regression_objective* class method), 346

TimeSeriesRegressionObjective (class in *evalml.pipelines*), 1080

is_multiclass() (in module *evalml.problem_types*), 1176

KNeighborsClassifier (class in *evalml.pipelines*), 1080

is_multiclass() (in module *evalml.problem_types.utils*), 1174

KNeighborsClassifier (class in *evalml.pipelines.components*), 907

is_regression() (in module *evalml.problem_types*), 1176

KNeighborsClassifier (class in *evalml.pipelines.components.estimators*), 612

is_regression() (in module *evalml.problem_types.utils*), 1174

KNeighborsClassifier (class in *evalml.pipelines.components.estimators.classifiers*), 492

is_search_space_exhausted() (*evalml.tuners.grid_search_tuner.GridSearchTuner* method), 1178

KNeighborsClassifier (class in *evalml.pipelines.components.estimators.classifiers.kneighbors_classifier*), 460

is_search_space_exhausted() (*evalml.tuners.GridSearchTuner* method), 1183

L

is_search_space_exhausted() (*evalml.tuners.random_search_tuner.RandomSearchTuner* method), 1179

LeadScoring (class in *evalml.objectives*), 381

LeadScoring (class in *evalml.objectives.lead_scoring*), 273

is_search_space_exhausted() (*evalml.tuners.RandomSearchTuner* method), 1184

LightGBMClassifier (class in *evalml.pipelines*), 1082

LightGBMClassifier (class in *evalml.pipelines.components*), 909

is_search_space_exhausted() (*evalml.tuners.skopt_tuner.SKOptTuner* method), 1180

LightGBMClassifier (class in *evalml.pipelines.components.estimators*), 615

is_search_space_exhausted() (*evalml.tuners.SKOptTuner* method), 1185

LightGBMClassifier (class in *evalml.pipelines.components.estimators.classifiers*), 495

is_search_space_exhausted() (*evalml.tuners.Tuner* method), 1186

LightGBMClassifier (class in *evalml.pipelines.components.estimators.classifiers.lightgbm_classifier*), 464

is_search_space_exhausted() (*evalml.tuners.tuner.Tuner* method), 1182

LightGBMRegressor (class in *evalml.pipelines*), 1085

is_time_series() (in module *evalml.problem_types*), 1176

LightGBMRegressor (class in *evalml.pipelines.components*), 913

is_time_series() (in module *evalml.problem_types.utils*), 1174

LightGBMRegressor (class in *evalml.pipelines.components.estimators*), 618

is_tree_estimator() (*evalml.model_family.model_family.ModelFamily* method), 232

is_tree_estimator()

[LightGBMRegressor](#) (class in [evalml.pipelines.components.estimators.regressors](#)), [1017](#)
[561](#) [linearized_component_graph\(\)](#)
[LightGBMRegressor](#) (class in [evalml.pipelines.components.estimators.regressors.lightgbm_regressor](#)), [1022](#)
[528](#) [linearized_component_graph\(\)](#)
[LinearDiscriminantAnalysis](#) (class in [evalml.pipelines.components](#)), [916](#) [\(evalml.pipelines.time_series_classification_pipelines.TimeSeries property\)](#), [1026](#)
[LinearDiscriminantAnalysis](#) (class in [evalml.pipelines.components.transformers](#)), [803](#) [linearized_component_graph\(\)](#) [\(evalml.pipelines.time_series_regression_pipeline.TimeSeriesReg property\)](#), [1031](#)
[LinearDiscriminantAnalysis](#) (class in [evalml.pipelines.components.transformers.dimensionality_reduction_pipeline](#)), [648](#) [linearized_component_graph\(\)](#) [\(evalml.pipelines.TimeSeriesBinaryClassificationPipeline property\)](#), [1139](#)
[LinearDiscriminantAnalysis](#) (class in [evalml.pipelines.components.transformers.dimensionality_reduction_pipeline](#)), [643](#) [linearized_component_graph\(\)](#) [\(evalml.pipelines.TimeSeriesClassificationPipeline property\)](#), [1143](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline property\)](#), [989](#) [linearized_component_graph\(\)](#) [\(evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property\)](#), [1147](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.BinaryClassificationPipeline property\)](#), [1040](#) [linearized_component_graph\(\)](#) [\(evalml.pipelines.TimeSeriesRegressionPipeline property\)](#), [1152](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.classification_pipeline.ClassificationPipeline property\)](#), [995](#) [LinearRegressor](#) (class in [evalml.pipelines](#)), [1088](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.ClassificationPipeline property\)](#), [1049](#) [LinearRegressor](#) (class in [evalml.pipelines.components.estimators](#)), [621](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.component_graph.ComponentGraph class method\)](#), [999](#) [LinearRegressor](#) (class in [evalml.pipelines.components.estimators.regressors](#)), [564](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.ComponentGraph class method\)](#), [1054](#) [LinearRegressor](#) (class in [evalml.pipelines.components.estimators.regressors.linear_regressor](#)), [531](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline property\)](#), [1003](#) [load\(\)](#) [\(evalml.automl.automl_search.AutoMLSearch static method\)](#), [179](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.MulticlassClassificationPipeline property\)](#), [1095](#) [load\(\)](#) [\(evalml.AutoMLSearch static method\)](#), [1201](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.pipeline_base.PipelineBase property\)](#), [1007](#) [load\(\)](#) [\(evalml.pipelines.ARIMARegressor static method\)](#), [1036](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.PipelineBase property\)](#), [1105](#) [load\(\)](#) [\(evalml.pipelines.binary_classification_pipeline.BinaryClassification static method\)](#), [989](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.regression_pipeline.RegressionPipeline property\)](#), [1013](#) [load\(\)](#) [\(evalml.pipelines.BinaryClassificationPipeline static method\)](#), [1040](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.RegressionPipeline property\)](#), [1113](#) [load\(\)](#) [\(evalml.pipelines.CatBoostClassifier static method\)](#), [1043](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.classification_pipeline.ClassificationPipeline static method\)](#), [995](#) [load\(\)](#) [\(evalml.pipelines.CatBoostRegressor static method\)](#), [1046](#)
[linearized_component_graph\(\)](#) [\(evalml.pipelines.ClassificationPipeline static method\)](#), [1049](#)

<code>load()</code> (<code>evalml.pipelines.components.ARIMARegressor</code> static method), 859	<code>load()</code> (<code>evalml.pipelines.components.estimators.CatBoostRegressor</code> static method), 592
<code>load()</code> (<code>evalml.pipelines.components.BaselineClassifier</code> static method), 861	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.baseline_classifier</code> static method), 447
<code>load()</code> (<code>evalml.pipelines.components.BaselineRegressor</code> static method), 864	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.BaselineClassifier</code> static method), 480
<code>load()</code> (<code>evalml.pipelines.components.CatBoostClassifier</code> static method), 866	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.catboost_classifier</code> static method), 450
<code>load()</code> (<code>evalml.pipelines.components.CatBoostRegressor</code> static method), 869	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.CatBoostClassifier</code> static method), 483
<code>load()</code> (<code>evalml.pipelines.components.component_base.ComponentBase</code> static method), 849	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.decision_tree_classifier</code> static method), 453
<code>load()</code> (<code>evalml.pipelines.components.ComponentBase</code> static method), 871	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier</code> static method), 486
<code>load()</code> (<code>evalml.pipelines.components.DateTimeFeaturizer</code> static method), 874	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.elasticnet_classifier</code> static method), 456
<code>load()</code> (<code>evalml.pipelines.components.DecisionTreeClassifier</code> static method), 876	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier</code> static method), 489
<code>load()</code> (<code>evalml.pipelines.components.DecisionTreeRegressor</code> static method), 879	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.et_classifier.ElasticNetClassifier</code> static method), 459
<code>load()</code> (<code>evalml.pipelines.components.DelayedFeatureTransformer</code> static method), 882	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier</code> static method), 492
<code>load()</code> (<code>evalml.pipelines.components.DFSTransformer</code> static method), 884	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.kneighbors_classifier</code> static method), 462
<code>load()</code> (<code>evalml.pipelines.components.DropColumns</code> static method), 886	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier</code> static method), 495
<code>load()</code> (<code>evalml.pipelines.components.DropNullColumns</code> static method), 888	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.lightgbm_classifier</code> static method), 466
<code>load()</code> (<code>evalml.pipelines.components.ElasticNetClassifier</code> static method), 891	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.LightGBMClassifier</code> static method), 498
<code>load()</code> (<code>evalml.pipelines.components.ElasticNetRegressor</code> static method), 893	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier</code> static method), 469
<code>load()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsembleBase</code> static method), 429	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier</code> static method), 501
<code>load()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsembleClassifier</code> static method), 432	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.RandomForestClassifier</code> static method), 503
<code>load()</code> (<code>evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsembleRegressor</code> static method), 435	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.rf_classifier.RandomForestClassifier</code> static method), 471
<code>load()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleBase</code> static method), 438	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.svm_classifier</code> static method), 474
<code>load()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleClassifier</code> static method), 441	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.SVMClassifier</code> static method), 506
<code>load()</code> (<code>evalml.pipelines.components.ensemble.StackedEnsembleRegressor</code> static method), 444	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.xgboost_classifier</code> static method), 477
<code>load()</code> (<code>evalml.pipelines.components.Estimator</code> static method), 895	<code>load()</code> (<code>evalml.pipelines.components.estimators.classifiers.XGBoostClassifier</code> static method), 508
<code>load()</code> (<code>evalml.pipelines.components.estimators.ARIMARegressor</code> static method), 582	<code>load()</code> (<code>evalml.pipelines.components.estimators.DecisionTreeClassifier</code> static method), 595
<code>load()</code> (<code>evalml.pipelines.components.estimators.BaselineClassifier</code> static method), 584	<code>load()</code> (<code>evalml.pipelines.components.estimators.DecisionTreeRegressor</code> static method), 598
<code>load()</code> (<code>evalml.pipelines.components.estimators.BaselineRegressor</code> static method), 587	<code>load()</code> (<code>evalml.pipelines.components.estimators.ElasticNetClassifier</code> static method), 601
<code>load()</code> (<code>evalml.pipelines.components.estimators.CatBoostClassifier</code> static method), 589	<code>load()</code> (<code>evalml.pipelines.components.estimators.ElasticNetRegressor</code> static method), 603

`load()` (`evalml.pipelines.components.estimators.Estimator` static method), 606

`load()` (`evalml.pipelines.components.estimators.estimator.ExtraTreesClassifier` static method), 578

`load()` (`evalml.pipelines.components.estimators.ExtraTreesClassifier` static method), 608

`load()` (`evalml.pipelines.components.estimators.ExtraTreesRegressor` static method), 611

`load()` (`evalml.pipelines.components.estimators.KNeighborsClassifier` static method), 614

`load()` (`evalml.pipelines.components.estimators.LightGBMClassifier` static method), 617

`load()` (`evalml.pipelines.components.estimators.LightGBMRegressor` static method), 620

`load()` (`evalml.pipelines.components.estimators.LinearRegressor` static method), 622

`load()` (`evalml.pipelines.components.estimators.LogisticRegressionClassifier` static method), 625

`load()` (`evalml.pipelines.components.estimators.RandomForestClassifier` static method), 627

`load()` (`evalml.pipelines.components.estimators.RandomForestRegressor` static method), 630

`load()` (`evalml.pipelines.components.estimators.regressor.ArimaRegressor` static method), 512

`load()` (`evalml.pipelines.components.estimators.regressor.ARMARegressor` static method), 547

`load()` (`evalml.pipelines.components.estimators.regressor.baseline_evalml_pipeline.BaselineRegressor` static method), 514

`load()` (`evalml.pipelines.components.estimators.regressor.BaselineRegressor` static method), 549

`load()` (`evalml.pipelines.components.estimators.regressor.catboost_regressor.CatBoostRegressor` static method), 517

`load()` (`evalml.pipelines.components.estimators.regressor.CatBoostRegressor` static method), 552

`load()` (`evalml.pipelines.components.estimators.regressor.decision_tree.DecisionTreeClassifier` static method), 521

`load()` (`evalml.pipelines.components.estimators.regressor.DecisionTreeRegressor` static method), 555

`load()` (`evalml.pipelines.components.estimators.regressor.elasticnet.ElasticNetRegressor` static method), 523

`load()` (`evalml.pipelines.components.estimators.regressor.ElasticNetRegressor` static method), 557

`load()` (`evalml.pipelines.components.estimators.regressor.linear_regression.LinearRegressor` static method), 527

`load()` (`evalml.pipelines.components.estimators.regressor.ExtraTreesRegressor` static method), 560

`load()` (`evalml.pipelines.components.estimators.regressor.lightgbm_regressor.LightGBMRegressor` static method), 530

`load()` (`evalml.pipelines.components.estimators.regressor.LightGBMRegressor` static method), 563

`load()` (`evalml.pipelines.components.estimators.regressor.linear_regressor.LinearRegressor` static method), 533

`load()` (`evalml.pipelines.components.estimators.regressor.LinearRegressor` static method), 566

`load()` (`evalml.pipelines.components.estimators.regressors.RandomForestClassifier` static method), 568

`load()` (`evalml.pipelines.components.estimators.regressors.rf_regressor.RFRegressor` static method), 535

`load()` (`evalml.pipelines.components.estimators.regressors.svm_regressor.SVMRegressor` static method), 538

`load()` (`evalml.pipelines.components.estimators.regressors.SVMRegressor` static method), 571

`load()` (`evalml.pipelines.components.estimators.regressors.time_series_baseline_evalml_pipeline.TimeSeriesBaselineRegressor` static method), 541

`load()` (`evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineRegressor` static method), 573

`load()` (`evalml.pipelines.components.estimators.regressors.xgboost_regressor.XGBoostRegressor` static method), 544

`load()` (`evalml.pipelines.components.estimators.regressors.XGBoostRegressor` static method), 576

`load()` (`evalml.pipelines.components.estimators.SVMClassifier` static method), 632

`load()` (`evalml.pipelines.components.estimators.SVMRegressor` static method), 634

`load()` (`evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator` static method), 637

`load()` (`evalml.pipelines.components.estimators.XGBoostClassifier` static method), 639

`load()` (`evalml.pipelines.components.estimators.XGBoostRegressor` static method), 642

`load()` (`evalml.pipelines.components.ExtraTreesClassifier` static method), 898

`load()` (`evalml.pipelines.components.ExtraTreesRegressor` static method), 901

`load()` (`evalml.pipelines.components.FeatureSelector` static method), 904

`load()` (`evalml.pipelines.components.Imputer` static method), 906

`load()` (`evalml.pipelines.components.DaemonTreeNeighborsClassifier` static method), 909

`load()` (`evalml.pipelines.components.LightGBMClassifier` static method), 912

`load()` (`evalml.pipelines.components.LightGBMRegressor` static method), 915

`load()` (`evalml.pipelines.components.LinearDiscriminantAnalysis` static method), 917

`load()` (`evalml.pipelines.components.LinearRegressor` static method), 919

`load()` (`evalml.pipelines.components.LogisticRegressionClassifier` static method), 922

`load()` (`evalml.pipelines.components.LSA` static method), 924

`load()` (`evalml.pipelines.components.OneHotEncoder` static method), 927

`load()` (`evalml.pipelines.components.PCA` static method), 929

`load()` (`evalml.pipelines.components.PerColumnImputer` static method), 931

<code>load()</code> (<code>evalml.pipelines.components.PolynomialDetrender</code> static method), 934	<code>load()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> static method), 649
<code>load()</code> (<code>evalml.pipelines.components.RandomForestClassifier</code> static method), 936	<code>load()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> static method), 652
<code>load()</code> (<code>evalml.pipelines.components.RandomForestRegressor</code> static method), 938	<code>load()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> static method), 647
<code>load()</code> (<code>evalml.pipelines.components.RFClassifierSelectFromModel</code> static method), 941	<code>load()</code> (<code>evalml.pipelines.components.transformers.DropColumns</code> static method), 795
<code>load()</code> (<code>evalml.pipelines.components.RFRegressorSelectFromModel</code> static method), 944	<code>load()</code> (<code>evalml.pipelines.components.transformers.DropNullColumns</code> static method), 797
<code>load()</code> (<code>evalml.pipelines.components.SelectColumns</code> static method), 946	<code>load()</code> (<code>evalml.pipelines.components.transformers.encoders.onehot_encoder</code> static method), 655
<code>load()</code> (<code>evalml.pipelines.components.SimpleImputer</code> static method), 948	<code>load()</code> (<code>evalml.pipelines.components.transformers.encoders.OneHotEncoder</code> static method), 662
<code>load()</code> (<code>evalml.pipelines.components.SMOTENCSampler</code> static method), 950	<code>load()</code> (<code>evalml.pipelines.components.transformers.encoders.target_encoder</code> static method), 659
<code>load()</code> (<code>evalml.pipelines.components.SMOTENSampler</code> static method), 953	<code>load()</code> (<code>evalml.pipelines.components.transformers.encoders.TargetEncoder</code> static method), 665
<code>load()</code> (<code>evalml.pipelines.components.SMOTESampler</code> static method), 955	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.feature_selector</code> static method), 668
<code>load()</code> (<code>evalml.pipelines.components.StackedEnsembleClassifier</code> static method), 957	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.FeatureSelector</code> static method), 677
<code>load()</code> (<code>evalml.pipelines.components.StackedEnsembleRegressor</code> static method), 960	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.rf_classifier</code> static method), 671
<code>load()</code> (<code>evalml.pipelines.components.StandardScaler</code> static method), 962	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.rf_regressor</code> static method), 674
<code>load()</code> (<code>evalml.pipelines.components.SVMClassifier</code> static method), 964	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.RFC</code> static method), 680
<code>load()</code> (<code>evalml.pipelines.components.SVMRegressor</code> static method), 966	<code>load()</code> (<code>evalml.pipelines.components.transformers.feature_selection.RFR</code> static method), 683
<code>load()</code> (<code>evalml.pipelines.components.TargetEncoder</code> static method), 969	<code>load()</code> (<code>evalml.pipelines.components.transformers.FeatureSelector</code> static method), 800
<code>load()</code> (<code>evalml.pipelines.components.TargetImputer</code> static method), 971	<code>load()</code> (<code>evalml.pipelines.components.transformers.Imputer</code> static method), 802
<code>load()</code> (<code>evalml.pipelines.components.TextFeaturizer</code> static method), 973	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.Imputer</code> static method), 697
<code>load()</code> (<code>evalml.pipelines.components.TimeSeriesBaselineEstimator</code> static method), 975	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.imputer.Imputer</code> static method), 686
<code>load()</code> (<code>evalml.pipelines.components.Transformer</code> static method), 978	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.per_column_imputer</code> static method), 688
<code>load()</code> (<code>evalml.pipelines.components.transformers.column_selector.ColumnSelector</code> static method), 776	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.PerColumnImputer</code> static method), 699
<code>load()</code> (<code>evalml.pipelines.components.transformers.column_selector.DropColumns</code> static method), 778	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.simple_imputer</code> static method), 691
<code>load()</code> (<code>evalml.pipelines.components.transformers.column_selector.SelectColumns</code> static method), 780	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.SimpleImputer</code> static method), 702
<code>load()</code> (<code>evalml.pipelines.components.transformers.DateTimeFeaturizer</code> static method), 789	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.target_imputer</code> static method), 694
<code>load()</code> (<code>evalml.pipelines.components.transformers.DelayedFeatureTransformer</code> static method), 791	<code>load()</code> (<code>evalml.pipelines.components.transformers.imputers.TargetImputer</code> static method), 704
<code>load()</code> (<code>evalml.pipelines.components.transformers.DFTransformer</code> static method), 793	<code>load()</code> (<code>evalml.pipelines.components.transformers.LinearDiscriminantAnalysis</code> static method), 804
<code>load()</code> (<code>evalml.pipelines.components.transformers.dimensionality_reducer</code> static method), 644	<code>load()</code> (<code>evalml.pipelines.components.transformers.LinearDiscriminantAnalysis</code> static method), 806

`load()` (`evalml.pipelines.components.transformers.OneHotEncoder` static method), 809
`load()` (`evalml.pipelines.components.transformers.PCA` static method), 811
`load()` (`evalml.pipelines.components.transformers.PerColumnImputer` static method), 814
`load()` (`evalml.pipelines.components.transformers.PolynomialDegree` static method), 816
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 707
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 728
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 710
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 730
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 733
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 712
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 735
`load()` (`evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer` static method), 715
`load()` (`evalml.pipelines.components.transformers.preprocessing.Lasso` static method), 737
`load()` (`evalml.pipelines.components.transformers.preprocessing.Lasso` static method), 717
`load()` (`evalml.pipelines.components.transformers.preprocessing.polynomial_degree` static method), 720
`load()` (`evalml.pipelines.components.transformers.preprocessing.PolynomialDegree` static method), 739
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextFeaturizer` static method), 722
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextFeaturizer` static method), 725
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextFeaturizer` static method), 741
`load()` (`evalml.pipelines.components.transformers.preprocessing.TextFeaturizer` static method), 744
`load()` (`evalml.pipelines.components.transformers.RFClassifier` static method), 819
`load()` (`evalml.pipelines.components.transformers.RFRegressor` static method), 822
`load()` (`evalml.pipelines.components.transformers.sampler.base` static method), 746
`load()` (`evalml.pipelines.components.transformers.sampler.base` static method), 749
`load()` (`evalml.pipelines.components.transformers.sampler.over` static method), 751
`load()` (`evalml.pipelines.components.transformers.sampler.over` static method), 754
`load()` (`evalml.pipelines.components.transformers.sampler.over` static method), 756
`load()` (`evalml.pipelines.components.transformers.samplers.SMOTE` static method), 762
`load()` (`evalml.pipelines.components.transformers.samplers.SMOTENS` static method), 764
`load()` (`evalml.pipelines.components.transformers.samplers.SMOTES` static method), 766
`load()` (`evalml.pipelines.components.transformers.samplers.Undersampler` static method), 769
`load()` (`evalml.pipelines.components.transformers.samplers.undersampler` static method), 759
`load()` (`evalml.pipelines.components.transformers.scalers.standard_scaler` static method), 771
`load()` (`evalml.pipelines.components.transformers.DelayedFeatureTransformer` static method), 774
`load()` (`evalml.pipelines.components.transformers.SelectColumns` static method), 824
`load()` (`evalml.pipelines.components.transformers.SimpleImputer` static method), 826
`load()` (`evalml.pipelines.components.transformers.sampler.SMOTE` static method), 828
`load()` (`evalml.pipelines.components.transformers.SMOTENS` static method), 831
`load()` (`evalml.pipelines.components.transformers.SMOTES` static method), 833
`load()` (`evalml.pipelines.components.transformers.StandardScaler` static method), 835
`load()` (`evalml.pipelines.components.transformers.TargetEncoder` static method), 837
`load()` (`evalml.pipelines.components.transformers.polynomial_degree` static method), 840
`load()` (`evalml.pipelines.components.transformers.TextFeaturizer` static method), 842
`load()` (`evalml.pipelines.components.transformers.TextFeaturizer` static method), 844
`load()` (`evalml.pipelines.components.transformers.transformer.TargetTransformer` static method), 783
`load()` (`evalml.pipelines.components.transformers.transformer.TargetTransformer` static method), 785
`load()` (`evalml.pipelines.components.transformers.Undersampler` static method), 847
`load()` (`evalml.pipelines.components.transformers.Undersampler` static method), 980
`load()` (`evalml.pipelines.components.XGBoostClassifier` static method), 983
`load()` (`evalml.pipelines.components.XGBoostRegressor` static method), 985
`load()` (`evalml.pipelines.components.DecisionTreeClassifier` static method), 1056
`load()` (`evalml.pipelines.components.DecisionTreeRegressor` static method), 1059
`load()` (`evalml.pipelines.components.DelayedFeatureTransformer` static method), 1061
`load()` (`evalml.pipelines.components.DelayedFeatureTransformer` static method), 1064

<code>load()</code> (<i>evalml.pipelines.ElasticNetClassifier static method</i>), 1066	<code>load()</code> (<i>evalml.pipelines.SVMClassifier static method</i>), 1131
<code>load()</code> (<i>evalml.pipelines.ElasticNetRegressor static method</i>), 1068	<code>load()</code> (<i>evalml.pipelines.SVMRegressor static method</i>), 1133
<code>load()</code> (<i>evalml.pipelines.Estimator static method</i>), 1071	<code>load()</code> (<i>evalml.pipelines.TargetEncoder static method</i>), 1136
<code>load()</code> (<i>evalml.pipelines.ExtraTreesClassifier static method</i>), 1074	<code>load()</code> (<i>evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassifier static method</i>), 1017
<code>load()</code> (<i>evalml.pipelines.ExtraTreesRegressor static method</i>), 1077	<code>load()</code> (<i>evalml.pipelines.time_series_classification_pipelines.TimeSeriesRegressor static method</i>), 1022
<code>load()</code> (<i>evalml.pipelines.FeatureSelector static method</i>), 1079	<code>load()</code> (<i>evalml.pipelines.time_series_classification_pipelines.TimeSeriesRegressor static method</i>), 1026
<code>load()</code> (<i>evalml.pipelines.KNeighborsClassifier static method</i>), 1082	<code>load()</code> (<i>evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline static method</i>), 1031
<code>load()</code> (<i>evalml.pipelines.LightGBMClassifier static method</i>), 1085	<code>load()</code> (<i>evalml.pipelines.TimeSeriesBinaryClassificationPipeline static method</i>), 1139
<code>load()</code> (<i>evalml.pipelines.LightGBMRegressor static method</i>), 1087	<code>load()</code> (<i>evalml.pipelines.TimeSeriesClassificationPipeline static method</i>), 1143
<code>load()</code> (<i>evalml.pipelines.LinearRegressor static method</i>), 1089	<code>load()</code> (<i>evalml.pipelines.TimeSeriesMulticlassClassificationPipeline static method</i>), 1148
<code>load()</code> (<i>evalml.pipelines.LogisticRegressionClassifier static method</i>), 1092	<code>load()</code> (<i>evalml.pipelines.TimeSeriesRegressionPipeline static method</i>), 1152
<code>load()</code> (<i>evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline static method</i>), 1003	<code>load()</code> (<i>evalml.pipelines.TimeSeriesRegressionPipeline static method</i>), 1154
<code>load()</code> (<i>evalml.pipelines.MulticlassClassificationPipeline static method</i>), 1096	<code>load()</code> (<i>evalml.pipelines.XGBoostClassifier static method</i>), 1157
<code>load()</code> (<i>evalml.pipelines.OneHotEncoder static method</i>), 1099	<code>load()</code> (<i>evalml.pipelines.XGBoostRegressor static method</i>), 1159
<code>load()</code> (<i>evalml.pipelines.PerColumnImputer static method</i>), 1101	<code>load_breast_cancer()</code> (<i>in module evalml.demos</i>), 228
<code>load()</code> (<i>evalml.pipelines.pipeline_base.PipelineBase static method</i>), 1007	<code>load_breast_cancer()</code> (<i>in module evalml.demos.breast_cancer</i>), 226
<code>load()</code> (<i>evalml.pipelines.PipelineBase static method</i>), 1105	<code>load_churn()</code> (<i>in module evalml.demos</i>), 228
<code>load()</code> (<i>evalml.pipelines.RandomForestClassifier static method</i>), 1107	<code>load_churn()</code> (<i>in module evalml.demos.churn</i>), 226
<code>load()</code> (<i>evalml.pipelines.RandomForestRegressor static method</i>), 1110	<code>load_data()</code> (<i>in module evalml.preprocessing</i>), 1169
<code>load()</code> (<i>evalml.pipelines.regression_pipeline.RegressionPipeline static method</i>), 1013	<code>load_data()</code> (<i>in module evalml.preprocessing.utils</i>), 1167
<code>load()</code> (<i>evalml.pipelines.RegressionPipeline static method</i>), 1113	<code>load_diabetes()</code> (<i>in module evalml.demos</i>), 228
<code>load()</code> (<i>evalml.pipelines.RFClassifierSelectFromModel static method</i>), 1116	<code>load_diabetes()</code> (<i>in module evalml.demos.diabetes</i>), 227
<code>load()</code> (<i>evalml.pipelines.RFRegressorSelectFromModel static method</i>), 1119	<code>load_fraud()</code> (<i>in module evalml.demos</i>), 229
<code>load()</code> (<i>evalml.pipelines.SimpleImputer static method</i>), 1121	<code>load_fraud()</code> (<i>in module evalml.demos.fraud</i>), 227
<code>load()</code> (<i>evalml.pipelines.StackedEnsembleClassifier static method</i>), 1124	<code>load_wine()</code> (<i>in module evalml.demos</i>), 229
<code>load()</code> (<i>evalml.pipelines.StackedEnsembleRegressor static method</i>), 1126	<code>load_wine()</code> (<i>in module evalml.demos.wine</i>), 228
<code>load()</code> (<i>evalml.pipelines.StandardScaler static method</i>), 1128	<code>log_error_callback()</code> (<i>in module evalml.automl.callbacks</i>), 172
	<code>log_subtitle()</code> (<i>in module evalml.utils</i>), 1197
	<code>log_subtitle()</code> (<i>in module evalml.utils.logger</i>), 1193
	<code>log_title()</code> (<i>in module evalml.utils</i>), 1197
	<code>log_title()</code> (<i>in module evalml.utils.logger</i>), 1193
	<code>logger</code> (<i>in module evalml.automl.automl_search</i>), 171
	<code>logger</code> (<i>in module evalml.automl.callbacks</i>), 172
	<code>logger</code> (<i>in module evalml.data_checks.no_variance_data_check</i>), 172

201
 logger (in module `evalml.objectives.sensitivity_low_alert`), 1033
 283
 logger (in module `evalml.pipelines.component_graph`), 1000
 logger (in module `evalml.pipelines.components.component_base`), 850
 logger (in module `evalml.pipelines.components.transformers.preprocessing`), 723
 logger (in module `evalml.pipelines.components.utils`), 852
 logger (in module `evalml.pipelines.pipeline_base`), 1005
 logger (in module `evalml.pipelines.utils`), 1033
 logger (in module `evalml.tuners.skopt_tuner`), 1180
 logger (in module `evalml.utils.cli_utils`), 1188
 logger (in module `evalml.utils.gen_utils`), 1191
 LogisticRegressionClassifier (class in `evalml.pipelines`), 1090
 LogisticRegressionClassifier (class in `evalml.pipelines.components`), 920
 LogisticRegressionClassifier (class in `evalml.pipelines.components.estimators`), 623
 LogisticRegressionClassifier (class in `evalml.pipelines.components.estimators.classifiers`), 499
 LogisticRegressionClassifier (class in `evalml.pipelines.components.estimators.classifiers.logistic_regression_classifier`), 467
 LogLossBinary (class in `evalml.objectives`), 383
 LogLossBinary (class in `evalml.objectives.standard_metrics`), 309
 LogLossMulticlass (class in `evalml.objectives`), 386
 LogLossMulticlass (class in `evalml.objectives.standard_metrics`), 311
 LSA (class in `evalml.pipelines.components`), 923
 LSA (class in `evalml.pipelines.components.transformers`), 805
 LSA (class in `evalml.pipelines.components.transformers.preprocessing`), 735
 LSA (class in `evalml.pipelines.components.transformers.preprocessing.lsa`), 716

M
 MAE (class in `evalml.objectives`), 387
 MAE (class in `evalml.objectives.standard_metrics`), 313
 make_balancing_dictionary() (in module `evalml.pipelines.components.utils`), 852
 make_data_splitter() (in module `evalml.automl`), 181
 make_data_splitter() (in module `evalml.automl.utils`), 174
 make_pipeline() (in module `evalml.pipelines.utils`), 1033
 MAPE (class in `evalml.objectives`), 389
 MAPE (class in `evalml.objectives.standard_metrics`), 315
 MaxError (class in `evalml.objectives`), 390
 MaxError (class in `evalml.objectives.standard_metrics`), 316
 MeanSquaredLogError (class in `evalml.objectives`), 392
 MCCBinary (class in `evalml.objectives.standard_metrics`), 318
 MCCMulticlass (class in `evalml.objectives`), 394
 MCCMulticlass (class in `evalml.objectives.standard_metrics`), 320
 MeanSquaredLogError (class in `evalml.objectives`), 396
 MeanSquaredLogError (class in `evalml.objectives.standard_metrics`), 321
 MedianAE (class in `evalml.objectives`), 397
 MedianAE (class in `evalml.objectives.standard_metrics`), 323
 method (in module `evalml.utils.update_checker`), 1193
 MethodPropertyNotFoundError, 229, 230
 MissingComponentError, 229, 230
 model_family() (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` property), 989
 model_family() (`evalml.pipelines.BinaryClassificationPipeline` property), 1040
 model_family() (`evalml.pipelines.classification_pipeline.ClassificationPipeline` property), 995
 model_family() (`evalml.pipelines.ClassificationPipeline` property), 1049
 model_family() (`evalml.pipelines.components.component_base.ComponentBase` property), 849
 model_family() (`evalml.pipelines.components.ComponentBase` property), 871
 model_family() (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` property), 1003
 model_family() (`evalml.pipelines.MulticlassClassificationPipeline` property), 1096
 model_family() (`evalml.pipelines.pipeline_base.PipelineBase` property), 1007
 model_family() (`evalml.pipelines.PipelineBase` property), 1105
 model_family() (`evalml.pipelines.regression_pipeline.RegressionPipeline` property), 1013
 model_family() (`evalml.pipelines.RegressionPipeline` property), 1113
 model_family() (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` property), 1018
 model_family() (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` property), 1022
 model_family() (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` property), 1026
 model_family() (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` property), 1018

[property](#)), 1031
[model_family\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#)
[property](#)), 1139
[model_family\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#)
[property](#)), 1143
[model_family\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#)
[property](#)), 1148
[model_family\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#)
[property](#)), 1152
[ModelFamily](#) (class in [evalml.model_family](#)), 233
[ModelFamily](#) (class in [evalml.model_family.model_family](#)), 231
[modifies_features\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#)
[property](#)), 849
[modifies_features\(\)](#) ([evalml.pipelines.components.ComponentBase](#)
[property](#)), 871
[modifies_target\(\)](#) ([evalml.pipelines.components.component_base.ComponentBase](#)
[property](#)), 849
[modifies_target\(\)](#) ([evalml.pipelines.components.ComponentBase](#)
[property](#)), 871
[module](#)
[evalml](#), 143
[evalml.automl](#), 143
[evalml.automl.automl_algorithm](#), 143
[evalml.automl.automl_algorithm.automl_algorithm](#), 144
[evalml.automl.automl_algorithm.iterative_automl_algorithm](#), 145
[evalml.automl.automl_search](#), 165
[evalml.automl.callbacks](#), 171
[evalml.automl.engine](#), 150
[evalml.automl.engine.cf_engine](#), 150
[evalml.automl.engine.dask_engine](#), 153
[evalml.automl.engine.engine_base](#), 155
[evalml.automl.engine.sequential_engine](#), 158
[evalml.automl.pipeline_search_plots](#), 172
[evalml.automl.utils](#), 173
[evalml.data_checks](#), 183
[evalml.data_checks.class_imbalance_data_check](#), 183
[evalml.data_checks.data_check](#), 185
[evalml.data_checks.data_check_action](#), 186
[evalml.data_checks.data_check_action_code](#), 187
[evalml.data_checks.data_check_message](#), 187
[evalml.data_checks.data_checks](#), 187
[evalml.data_checks.data_check_message_code](#), 189
[evalml.data_checks.data_check_message_type](#), 191
[evalml.data_checks.datetime_nan_data_check](#), 191
[evalml.data_checks.default_data_checks](#), 193
[evalml.data_checks.highly_null_data_check](#), 195
[evalml.data_checks.id_columns_data_check](#), 196
[evalml.data_checks.invalid_targets_data_check](#), 197
[evalml.data_checks.multicollinearity_data_check](#), 199
[evalml.data_checks.natural_language_nan_data_check](#), 200
[evalml.data_checks.no_variance_data_check](#), 201
[evalml.data_checks.outliers_data_check](#), 202
[evalml.data_checks.sparsity_data_check](#), 203
[evalml.data_checks.target_leakage_data_check](#), 204
[evalml.data_checks.uniqueness_data_check](#), 206
[evalml.data_checks.utils](#), 207
[evalml.demos](#), 226
[evalml.demos.breast_cancer](#), 226
[evalml.demos.churn](#), 226
[evalml.demos.diabetes](#), 227
[evalml.demos.fraud](#), 227
[evalml.demos.wine](#), 228
[evalml.exceptions](#), 229
[evalml.exceptions.exceptions](#), 229
[evalml.model_family](#), 231
[evalml.model_family.model_family](#), 231
[evalml.model_family.utils](#), 232
[evalml.model_understanding](#), 234
[evalml.model_understanding.force_plots](#), 241
[evalml.model_understanding.graphs](#), 242
[evalml.model_understanding.permutation_importance](#), 252
[evalml.model_understanding.prediction_explanations](#), 252
[evalml.model_understanding.prediction_explanations.force_plots](#), 253
[evalml.model_understanding.prediction_explanations.graphs](#), 254

<code>evalml.objectives</code> , 265	<code>evalml.pipelines.components.estimators.classifi</code>
<code>evalml.objectives.binary_classification_objective</code> , 265	<code>evalml.pipelines.components.estimators.classifi</code>
<code>evalml.objectives.cost_benefit_matrix</code> , 268	460
<code>evalml.objectives.fraud_cost</code> , 270	463
<code>evalml.objectives.lead_scoring</code> , 273	<code>evalml.pipelines.components.estimators.classifi</code>
<code>evalml.objectives.multiclass_classification_objective</code> , 276	<code>evalml.pipelines.components.estimators.classifi</code>
<code>evalml.objectives.objective_base</code> , 278	470
<code>evalml.objectives.regression_objective</code> , 280	472
<code>evalml.objectives.sensitivity_low_alert</code> , 283	475
<code>evalml.objectives.standard_metrics</code> , 285	577
<code>evalml.objectives.time_series_regression_objective</code> , 345	500
<code>evalml.objectives.utils</code> , 347	509
<code>evalml.pipelines</code> , 427	<code>evalml.pipelines.components.estimators.regresso</code>
<code>evalml.pipelines.binary_classification_pipeline</code> , 486	<code>evalml.pipelines.components.estimators.regresso</code>
<code>evalml.pipelines.binary_classification_pipeline_base</code> , 491	<code>evalml.pipelines.components.estimators.regresso</code>
<code>evalml.pipelines.classification_pipeline</code> , 492	518
<code>evalml.pipelines.component_graph</code> , 496	522
<code>evalml.pipelines.components</code> , 427	524
<code>evalml.pipelines.components.component_base</code> , 448	528
<code>evalml.pipelines.components.component_base_abstract</code> , 450	531
<code>evalml.pipelines.components.ensemble</code> , 427	<code>evalml.pipelines.components.estimators.regresso</code>
<code>evalml.pipelines.components.ensemble.stacked_ensemble_base</code> , 427	536
<code>evalml.pipelines.components.ensemble.stacked_ensemble_classification</code> , 430	539
<code>evalml.pipelines.components.ensemble.stacked_ensemble_regression</code> , 433	542
<code>evalml.pipelines.components.estimators</code> , 445	<code>evalml.pipelines.components.transformers</code> , 643
<code>evalml.pipelines.components.estimators.classification</code> , 445	774
<code>evalml.pipelines.components.estimators.classification_base</code> , 445	643
<code>evalml.pipelines.components.estimators.classification.boosted</code> , 448	643
<code>evalml.pipelines.components.estimators.classification.decision_tree</code> , 451	645
<code>evalml.pipelines.components.estimators.classification.elasticnet</code> , 454	652

[evalml.pipelines.components.transformers.sequential_pipeline_components_utils](#),
 652
[evalml.pipelines.components.transformers.sequential_pipeline_components_utils](#),
 657
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1004
 666
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1009
 666
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1010
 668
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1014
 672
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1028
 683
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1032
 683
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1160
 683
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1160
 686
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1160
 689
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1161
 692
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1162
 705
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1163
 705
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1167
 708
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1172
 710
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1173
 713
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1177
 715
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1178
 718
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1180
 721
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1181
 723
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1182
 744
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1187
 744
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1189
 744
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1192
 744
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1193
 749
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1193
 757
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1193
 770
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1193
 770
[evalml.pipelines.components.transformers.sequential_pipeline_base](#), 1193
 781

`evalml.pipelines.multiclass_classification_pipeline` (class in `evalml.pipelines`), 1000

`MulticollinearityDataCheck` (class in `evalml.data_checks`), 219

`MulticollinearityDataCheck` (class in `evalml.data_checks.multicollinearity_data_check`), 199

N

`name()` (`evalml.data_checks.class_imbalance_data_check.ClassImbalanceDataCheck` method), 184

`name()` (`evalml.data_checks.ClassImbalanceDataCheck` method), 209

`name()` (`evalml.data_checks.data_check.DataCheck` method), 186

`name()` (`evalml.data_checks.data_check_action_code.DataCheckActionCode` method), 187

`name()` (`evalml.data_checks.data_check_message_code.DataCheckMessageCode` method), 190

`name()` (`evalml.data_checks.data_check_message_type.DataCheckMessageType` method), 191

`name()` (`evalml.data_checks.DataCheck` method), 211

`name()` (`evalml.data_checks.DataCheckActionCode` method), 211

`name()` (`evalml.data_checks.DataCheckMessageCode` method), 213

`name()` (`evalml.data_checks.DataCheckMessageType` method), 214

`name()` (`evalml.data_checks.datetime_nan_data_check.DateTimeNaNDataCheck` method), 192

`name()` (`evalml.data_checks.DateTImeNaNDataCheck` method), 215

`name()` (`evalml.data_checks.highly_null_data_check.HighlyNullDataCheck` method), 195

`name()` (`evalml.data_checks.HighlyNullDataCheck` method), 217

`name()` (`evalml.data_checks.id_columns_data_check.IDColumnsDataCheck` method), 196

`name()` (`evalml.data_checks.IDColumnsDataCheck` method), 218

`name()` (`evalml.data_checks.invalid_targets_data_check.InvalidTargetDataCheck` method), 198

`name()` (`evalml.data_checks.InvalidTargetDataCheck` method), 219

`name()` (`evalml.data_checks.multicollinearity_data_check.MulticollinearityDataCheck` method), 199

`name()` (`evalml.data_checks.MulticollinearityDataCheck` method), 220

`name()` (`evalml.data_checks.natural_language_nan_data_check.NaturalLanguageNaNDataCheck` method), 200

`name()` (`evalml.data_checks.NaturalLanguageNaNDataCheck` method), 220

`name()` (`evalml.data_checks.no_variance_data_check.NoVarianceDataCheck` method), 201

`name()` (`evalml.data_checks.NoVarianceDataCheck` method), 221

`name()` (`evalml.data_checks.outliers_data_check.OutliersDataCheck` method), 202

`name()` (`evalml.data_checks.OutliersDataCheck` method), 221

`name()` (`evalml.data_checks.sparsity_data_check.SparsityDataCheck` method), 203

`name()` (`evalml.data_checks.SparsityDataCheck` method), 220

`name()` (`evalml.data_checks.target_leakage_data_check.TargetLeakageDataCheck` method), 205

`name()` (`evalml.data_checks.TargetLeakageDataCheck` method), 224

`name()` (`evalml.data_checks.uniqueness_data_check.UniquenessDataCheck` method), 206

`name()` (`evalml.data_checks.UniquenessDataCheck` method), 225

`name()` (`evalml.model_family.model_family.ModelFamily` method), 232

`name()` (`evalml.model_family.ModelFamily` method), 234

`name()` (`evalml.model_understanding.prediction_explanations.explainers` method), 238

`name()` (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` property), 266

`name()` (`evalml.objectives.BinaryClassificationObjective` property), 366

`name()` (`evalml.objectives.datetime_nan_data_check.DateTimeNaNDataCheck` property), 277

`name()` (`evalml.objectives.MulticlassClassificationObjective` property), 401

`name()` (`evalml.objectives.objective_base.ObjectiveBase` property), 279

`name()` (`evalml.objectives.ObjectiveBase` property), 403

`name()` (`evalml.objectives.regression_objective.RegressionObjective` property), 421

`name()` (`evalml.objectives.RegressionObjective` property), 421

`name()` (`evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective` property), 426

`name()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` property), 989

`name()` (`evalml.pipelines.BinaryClassificationPipeline` property), 990

`name()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` property), 995

`name()` (`evalml.pipelines.ClassificationPipeline` property), 995

`name()` (`evalml.pipelines.natural_language_nan_data_check.NaturalLanguageNaNDataCheck` property), 850

`name()` (`evalml.pipelines.components.component_base.ComponentBase` property), 871

`name()` (`evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase` property), 871

property), 430

name () (evalml.pipelines.components.ensemble.StackedEnsembleClassifier property), 438

name () (evalml.pipelines.components.Estimator property), 896

name () (evalml.pipelines.components.estimators.Estimator property), 606

name () (evalml.pipelines.components.estimators.estimator.Estimator property), 579

name () (evalml.pipelines.components.FeatureSelector property), 904

name () (evalml.pipelines.components.Transformer property), 978

name () (evalml.pipelines.components.transformers.column_selector.ColumnSelector property), 776

name () (evalml.pipelines.components.transformers.feature_selection.feature_selector.FeatureSelector property), 668

name () (evalml.pipelines.components.transformers.feature_selection.feature_selector.FeatureSelector property), 677

name () (evalml.pipelines.components.transformers.FeatureSelector property), 800

name () (evalml.pipelines.components.transformers.preprocessing_text_transformer.TextTransformer property), 725

name () (evalml.pipelines.components.transformers.preprocessing_text_transformer.TextTransformer property), 744

name () (evalml.pipelines.components.transformers.sampler_base_sampler.BaseSampler property), 747

name () (evalml.pipelines.components.transformers.sampler_base_sampler.BaseSampler property), 749

name () (evalml.pipelines.components.transformers.Transformers_fitting() method), 844

name () (evalml.pipelines.components.transformers.transformer_base_transformer.BaseTransformer property), 783

name () (evalml.pipelines.components.transformers.transformer_base_transformer.BaseTransformer property), 785

name () (evalml.pipelines.Estimator property), 1071

name () (evalml.pipelines.FeatureSelector property), 1079

name () (evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline property), 1003

name () (evalml.pipelines.MulticlassClassificationPipeline property), 1096

name () (evalml.pipelines.pipeline_base.PipelineBase property), 1007

name () (evalml.pipelines.PipelineBase property), 1105

name () (evalml.pipelines.regression_pipeline.RegressionPipeline property), 1013

name () (evalml.pipelines.RegressionPipeline property), 1113

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1018

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1022

name () (evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline property), 1026

name () (evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline property), 1031

name () (evalml.pipelines.TimeSeriesBinaryClassificationPipeline property), 1139

name () (evalml.pipelines.TimeSeriesClassificationPipeline property), 1143

name () (evalml.pipelines.TimeSeriesMulticlassClassificationPipeline property), 1148

name () (evalml.pipelines.TimeSeriesRegressionPipeline property), 1152

name () (evalml.pipelines.Transformer property), 1155

name () (evalml.problem_types.problem_types.ProblemTypes method), 220

name () (evalml.problem_types.ProblemTypes method), NaturalLanguageNaNDataCheck (class in evalml.data_checks.natural_language_nan_data_check), 200

name () (evalml.pipelines.components.ARIMARegressor method), 1037

name () (evalml.pipelines.CatBoostClassifier method), 1044

name () (evalml.pipelines.CatBoostRegressor method), 1046

name () (evalml.pipelines.components.ARIMARegressor method), 859

needs_fitting() (evalml.pipelines.components.BaselineClassifier method), 862

needs_fitting() (evalml.pipelines.components.BaselineRegressor method), 864

needs_fitting() (evalml.pipelines.components.CatBoostClassifier method), 866

needs_fitting() (evalml.pipelines.components.CatBoostRegressor method), 869

needs_fitting() (evalml.pipelines.components.component_base.ComponentBase method), 871

needs_fitting() (evalml.pipelines.components.ComponentBase method), 874

needs_fitting() (evalml.pipelines.components.DateTimeFeaturizer method), 877

needs_fitting() (evalml.pipelines.components.DecisionTreeClassifier method), 880

needs_fitting() (evalml.pipelines.components.DFSTransformer method), 884

needs_fitting() (evalml.pipelines.components.DropNullColumns method), 888

needs_fitting() (evalml.pipelines.components.ElasticNetClassifier method), 891

needs_fitting() (evalml.pipelines.components.ElasticNetRegressor method), 891

method), 893

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble), 430

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble_classifier), 433

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble_regressor), 436

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble_base_classifier), 439

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble_classifier), 441

needs_fitting() (evalml.pipelines.components.ensemble_stacked_ensemble_regressor), 444

needs_fitting() (evalml.pipelines.components.Estimator), 896

needs_fitting() (evalml.pipelines.components.estimator.ARIMARegressor), 582

needs_fitting() (evalml.pipelines.components.estimator.BaseLineClassifier), 585

needs_fitting() (evalml.pipelines.components.estimator.BaseLineRegressor), 587

needs_fitting() (evalml.pipelines.components.estimator.GradientBoostClassifier), 590

needs_fitting() (evalml.pipelines.components.estimator.GradientBoostRegressor), 592

needs_fitting() (evalml.pipelines.components.estimator.classifier_base_line_classifier), 447

needs_fitting() (evalml.pipelines.components.estimator.classifier_base_line_classifier), 480

needs_fitting() (evalml.pipelines.components.estimator.classifier_catboost_classifier), 450

needs_fitting() (evalml.pipelines.components.estimator.classifier_catboost_classifier), 483

needs_fitting() (evalml.pipelines.components.estimator.classifier_decision_tree_classifier), 453

needs_fitting() (evalml.pipelines.components.estimator.classifier_decision_tree_classifier), 486

needs_fitting() (evalml.pipelines.components.estimator.classifier_elasticnet_classifier), 456

needs_fitting() (evalml.pipelines.components.estimator.classifier_elasticnet_classifier), 489

needs_fitting() (evalml.pipelines.components.estimator.classifier_extra_trees_classifier), 459

needs_fitting() (evalml.pipelines.components.estimator.classifier_extra_trees_classifier), 492

needs_fitting() (evalml.pipelines.components.estimator.classifier_kneighbors_classifier), 462

needs_fitting() (evalml.pipelines.components.estimator.classifier_kneighbors_classifier), 495

needs_fitting() (evalml.pipelines.components.estimator.classifier_lightgbm_classifier), 466

needs_fitting() (evalml.pipelines.components.estimator.classifier_lightgbm_classifier), 498

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 469

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 501

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 503

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 472

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 474

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 506

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 478

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 508

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 595

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 598

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 601

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 603

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 606

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 579

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 609

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 611

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 614

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 617

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 620

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 622

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 625

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 627

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 630

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 512

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 547

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 515

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 549

needs_fitting() (evalml.pipelines.components.estimator.classifier_logistic_regression_classifier), 549

method), 518

needs_fitting() (evalml.pipelines.components.estimators.regressors.CatBoostRegressor), 552

needs_fitting() (evalml.pipelines.components.estimators.regressors.decision_tree.DecisionTreeRegressor), 521

needs_fitting() (evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor), 555

needs_fitting() (evalml.pipelines.components.estimators.regressors.elasticnet.ElasticNetRegressor), 524

needs_fitting() (evalml.pipelines.components.estimators.regressors.ElasticNetRegressor), 558

needs_fitting() (evalml.pipelines.components.estimators.regressors.linear.LinearRegressor), 527

needs_fitting() (evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor), 560

needs_fitting() (evalml.pipelines.components.estimators.regressors.lightgbm.LightGBMRegressor), 530

needs_fitting() (evalml.pipelines.components.estimators.regressors.LightGBMRegressor), 563

needs_fitting() (evalml.pipelines.components.estimators.regressors.linear.LinearRegressor), 533

needs_fitting() (evalml.pipelines.components.estimators.regressors.LinearRegressor), 566

needs_fitting() (evalml.pipelines.components.estimators.regressors.RandomForestRegressor), 568

needs_fitting() (evalml.pipelines.components.estimators.regressors.iforest.IforestRegressor), 536

needs_fitting() (evalml.pipelines.components.estimators.regressors.svm.SVRRegressor), 538

needs_fitting() (evalml.pipelines.components.estimators.regressors.SVRRegressor), 571

needs_fitting() (evalml.pipelines.components.estimators.regressors.time_series.TimeSeriesRegressor), 541

needs_fitting() (evalml.pipelines.components.estimators.regressors.TimeSeriesRegressor), 573

needs_fitting() (evalml.pipelines.components.estimators.regressors.xgboost.XGBRegressor), 544

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 576

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 632

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 635

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 637

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 639

needs_fitting() (evalml.pipelines.components.estimators.regressors.XGBRegressor), 642

needs_fitting() (evalml.pipelines.components.ExtraTreesClassifier), 899

needs_fitting() (evalml.pipelines.components.ExtraTreesRegressor), 901

needs_fitting() (evalml.pipelines.components.FeatureSelector), 904

needs_fitting() (evalml.pipelines.components.Imputer), 906

needs_fitting() (evalml.pipelines.components.neighbors.KNeighborsClassifier), 909

needs_fitting() (evalml.pipelines.components.LightGBMClassifier), 912

needs_fitting() (evalml.pipelines.components.LightGBMRegressor), 915

needs_fitting() (evalml.pipelines.components.LinearDiscriminantAnalysis), 917

needs_fitting() (evalml.pipelines.components.LinearRegressor), 919

needs_fitting() (evalml.pipelines.components.LogisticRegressionClassifier), 922

needs_fitting() (evalml.pipelines.components.LogisticRegressionClassifier), 924

needs_fitting() (evalml.pipelines.components.OneHotEncoder), 927

needs_fitting() (evalml.pipelines.components.PCA), 929

needs_fitting() (evalml.pipelines.components.PerColumnImputer), 932

needs_fitting() (evalml.pipelines.components.PolynomialDetrender), 934

needs_fitting() (evalml.pipelines.components.RandomForestClassifier), 936

needs_fitting() (evalml.pipelines.components.RandomForestRegressor), 939

needs_fitting() (evalml.pipelines.components.RFClassifierSelectFromModel), 941

needs_fitting() (evalml.pipelines.components.SBFBRegressorSelectFromModel), 944

needs_fitting() (evalml.pipelines.components.SimpleImputer), 948

needs_fitting() (evalml.pipelines.components.SMOTEKerasRegressor), 951

needs_fitting() (evalml.pipelines.components.SMOTENSampler), 953

needs_fitting() (evalml.pipelines.components.SMOTENSampler), 955

needs_fitting() (evalml.pipelines.components.StackedEnsembleClassifier), 958

needs_fitting() (evalml.pipelines.components.StackedEnsembleRegressor), 960

needs_fitting() (evalml.pipelines.components.StandardScaler), 962

needs_fitting() (evalml.pipelines.components.SVMClassifier), 964

needs_fitting() (evalml.pipelines.components.SVMRegressor), 967

needs_fitting() (evalml.pipelines.components.TargetEncoder), 969

needs_fitting() (evalml.pipelines.components.TargetImputer), 971

method), 971

needs_fitting() (evalml.pipelines.components.TextFeaturizer), 973

needs_fitting() (evalml.pipelines.components.TimeSeriesBaseLineEstimator), 976

needs_fitting() (evalml.pipelines.components.Transformers), 978

needs_fitting() (evalml.pipelines.components.transformers.ColumnSelector), 776

needs_fitting() (evalml.pipelines.components.transformers.DateTimeFeaturizer), 789

needs_fitting() (evalml.pipelines.components.transformers.DFSTransformer), 793

needs_fitting() (evalml.pipelines.components.transformers.DimensionalityReducer), 645

needs_fitting() (evalml.pipelines.components.transformers.DimensionalityReducer), 649

needs_fitting() (evalml.pipelines.components.transformers.DimensionalityReducer), 652

needs_fitting() (evalml.pipelines.components.transformers.DimensionalityReducer), 647

needs_fitting() (evalml.pipelines.components.transformers.DropNullColumns), 797

needs_fitting() (evalml.pipelines.components.transformers.Encoders), 655

needs_fitting() (evalml.pipelines.components.transformers.Encoders), 662

needs_fitting() (evalml.pipelines.components.transformers.Encoders), 659

needs_fitting() (evalml.pipelines.components.transformers.Encoders), 665

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 668

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 677

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 671

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 674

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 680

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 683

needs_fitting() (evalml.pipelines.components.transformers.FeatureSelector), 800

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 802

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 697

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 686

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 688

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 699

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 691

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 702

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 694

needs_fitting() (evalml.pipelines.components.transformers.Imputer), 704

needs_fitting() (evalml.pipelines.components.transformers.LinearDiscriminantAnalysis), 804

needs_fitting() (evalml.pipelines.components.transformers.LSA), 806

needs_fitting() (evalml.pipelines.components.transformers.L2Lasso), 809

needs_fitting() (evalml.pipelines.components.transformers.L2Lasso), 812

needs_fitting() (evalml.pipelines.components.transformers.L2Lasso), 814

needs_fitting() (evalml.pipelines.components.transformers.L2Lasso), 816

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 707

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 728

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 733

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 712

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 735

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 715

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 737

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 717

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 720

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 739

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 722

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 725

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 742

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 744

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 819

needs_fitting() (evalml.pipelines.components.transformers.Preprocessor), 822

needs_fitting() (evalml.pipelines.components.transformers.Samplers), 822

method), 747

needs_fitting() (evalml.pipelines.components.transformers.samplers.base), 749

needs_fitting() (evalml.pipelines.components.transformers.samplers.over), 751

needs_fitting() (evalml.pipelines.components.transformers.samplers.over), 754

needs_fitting() (evalml.pipelines.components.transformers.samplers.over), 756

needs_fitting() (evalml.pipelines.components.transformers.samplers.SMOTE), 762

needs_fitting() (evalml.pipelines.components.transformers.samplers.SMOTE), 764

needs_fitting() (evalml.pipelines.components.transformers.samplers.SMOTE), 766

needs_fitting() (evalml.pipelines.components.transformers.samplers.undersample), 769

needs_fitting() (evalml.pipelines.components.transformers.samplers.undersample), 759

needs_fitting() (evalml.pipelines.components.transformers.scaler.standard), 771

needs_fitting() (evalml.pipelines.components.transformers.scaler.standard), 774

needs_fitting() (evalml.pipelines.components.transformers.SimpleImputer), 826

needs_fitting() (evalml.pipelines.components.transformers.SMOTE), 828

needs_fitting() (evalml.pipelines.components.transformers.SMOTE), 831

needs_fitting() (evalml.pipelines.components.transformers.SMOTE), 833

needs_fitting() (evalml.pipelines.components.transformers.StandardScaler), 835

needs_fitting() (evalml.pipelines.components.transformers.TargetEncoder), 838

needs_fitting() (evalml.pipelines.components.transformers.TargetImputer), 840

needs_fitting() (evalml.pipelines.components.transformers.TargetImputer), 842

needs_fitting() (evalml.pipelines.components.transformers.Transformer), 844

needs_fitting() (evalml.pipelines.components.transformers.transformer.Target), 783

needs_fitting() (evalml.pipelines.components.transformers.transformer.Target), 785

needs_fitting() (evalml.pipelines.components.transformers.undersampler), 847

needs_fitting() (evalml.pipelines.components.undersampler), 980

needs_fitting() (evalml.pipelines.components.XGBoostClassifier), 983

needs_fitting() (evalml.pipelines.components.XGBoostRegressor), 985

needs_fitting() (evalml.pipelines.DecisionTreeClassifier), 1056

needs_fitting() (evalml.pipelines.DecisionTreeRegressor), 1059

needs_fitting() (evalml.pipelines.estimators.GradientBoostingClassifier), 1064

needs_fitting() (evalml.pipelines.estimators.GradientBoostingRegressor), 1066

needs_fitting() (evalml.pipelines.estimators.GradientBoostingRegressor), 1069

needs_fitting() (evalml.pipelines.estimators.GradientBoostingRegressor), 1071

needs_fitting() (evalml.pipelines.ExtraTreesClassifier), 1074

needs_fitting() (evalml.pipelines.ExtraTreesRegressor), 1077

needs_fitting() (evalml.pipelines.FeatureSelector), 1079

needs_fitting() (evalml.pipelines.KNeighborsClassifier), 1082

needs_fitting() (evalml.pipelines.LightGBMClassifier), 1085

needs_fitting() (evalml.pipelines.LightGBMRegressor), 1087

needs_fitting() (evalml.pipelines.LinearRegressor), 1090

needs_fitting() (evalml.pipelines.LogisticRegressionClassifier), 1092

needs_fitting() (evalml.pipelines.OneHotEncoder), 1099

needs_fitting() (evalml.pipelines.PerColumnImputer), 1102

needs_fitting() (evalml.pipelines.RandomForestClassifier), 1108

needs_fitting() (evalml.pipelines.RandomForestRegressor), 1110

needs_fitting() (evalml.pipelines.RFClassifierSelectFromModel), 1116

needs_fitting() (evalml.pipelines.RFRegressorSelectFromModel), 1119

needs_fitting() (evalml.pipelines.SimpleImputer), 1121

needs_fitting() (evalml.pipelines.StackedEnsembleClassifier), 1124

needs_fitting() (evalml.pipelines.StackedEnsembleRegressor), 1126

needs_fitting() (evalml.pipelines.StandardScaler), 1128

needs_fitting() (evalml.pipelines.SVMClassifier), 1131

needs_fitting() (evalml.pipelines.SVMRegressor), 1133

needs_fitting() (evalml.pipelines.TargetEncoder), 1136

needs_fitting() (evalml.pipelines.Transformer), 1136

[method](#)), 1155
[needs_fitting\(\)](#) ([evalml.pipelines.XGBoostClassifier](#) [method](#)), 1157
[needs_fitting\(\)](#) ([evalml.pipelines.XGBoostRegressor](#) [method](#)), 1159
[new\(\)](#) ([evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline](#) [method](#)), 989
[new\(\)](#) ([evalml.pipelines.BinaryClassificationPipeline](#) [method](#)), 1040
[new\(\)](#) ([evalml.pipelines.classification_pipeline.ClassificationPipeline](#) [method](#)), 995
[new\(\)](#) ([evalml.pipelines.ClassificationPipeline](#) [method](#)), 1050
[new\(\)](#) ([evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline](#) [method](#)), 1003
[new\(\)](#) ([evalml.pipelines.MulticlassClassificationPipeline](#) [method](#)), 1096
[new\(\)](#) ([evalml.pipelines.pipeline_base.PipelineBase](#) [method](#)), 1008
[new\(\)](#) ([evalml.pipelines.PipelineBase](#) [method](#)), 1105
[new\(\)](#) ([evalml.pipelines.regression_pipeline.RegressionPipeline](#) [method](#)), 1013
[new\(\)](#) ([evalml.pipelines.RegressionPipeline](#) [method](#)), 1113
[new\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesBinaryClassificationPipeline](#) [method](#)), 1018
[new\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline](#) [method](#)), 1022
[new\(\)](#) ([evalml.pipelines.time_series_classification_pipelines.TimeSeriesMulticlassClassificationPipeline](#) [method](#)), 1027
[new\(\)](#) ([evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline](#) [method](#)), 1031
[new\(\)](#) ([evalml.pipelines.TimeSeriesBinaryClassificationPipeline](#) [method](#)), 1139
[new\(\)](#) ([evalml.pipelines.TimeSeriesClassificationPipeline](#) [method](#)), 1144
[new\(\)](#) ([evalml.pipelines.TimeSeriesMulticlassClassificationPipeline](#) [method](#)), 1148
[new\(\)](#) ([evalml.pipelines.TimeSeriesRegressionPipeline](#) [method](#)), 1152
[next_batch\(\)](#) ([evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm](#) [method](#)), 145
[next_batch\(\)](#) ([evalml.automl.automl_algorithm.AutoMLAlgorithm](#) [method](#)), 149
[next_batch\(\)](#) ([evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm](#) [method](#)), 147
[next_batch\(\)](#) ([evalml.automl.automl_algorithm.IterativeAlgorithm](#) [method](#)), 150
[NoParamsException](#), 1182, 1184
[NoPositiveLabelException](#), 229, 230
[normalize_confusion_matrix\(\)](#) (in module [evalml.model_understanding](#)), 262
[normalize_confusion_matrix\(\)](#) (in module [evalml.model_understanding.graphs](#)), 248
[NoVarianceDataCheck](#) (class in [evalml.data_checks](#)), 221
[NoVarianceDataCheck](#) (class in [evalml.data_checks.no_variance_data_check](#)), 201
[number_of_features\(\)](#) (in module [evalml.preprocessing](#)), 1169
[number_of_features\(\)](#) (in module [evalml.preprocessing.utils](#)), 1168
[numeric_and_boolean_wv](#) (in module [evalml.utils.woodwork_utils](#)), 1194
[objective_function\(\)](#) ([evalml.objectives.AccuracyBinary](#) [method](#)), 351
[objective_function\(\)](#) ([evalml.objectives.AccuracyMulticlass](#) [method](#)), 353
[objective_function\(\)](#) ([evalml.objectives.AUC](#) [method](#)), 355
[objective_function\(\)](#) ([evalml.objectives.AUCMacro](#) [method](#)), 357
[objective_function\(\)](#) ([evalml.objectives.AUCMicro](#) [method](#)), 358
[objective_function\(\)](#) ([evalml.objectives.AUCWeighted](#) [method](#)), 360
[objective_function\(\)](#) ([evalml.objectives.BalancedAccuracyBinary](#) [method](#)), 362
[objective_function\(\)](#) ([evalml.objectives.BalancedAccuracyMulticlass](#) [method](#)), 364
[objective_function\(\)](#) ([evalml.objectives.binary_classification_objective.BinaryClassificationObjective](#) [class method](#)), 266
[objective_function\(\)](#) ([evalml.objectives.BinaryClassificationObjective](#) [class method](#)), 300
[objective_function\(\)](#) ([evalml.objectives.cost_benefit_matrix.CostBenefitMatrix](#) [method](#)), 269
[objective_function\(\)](#) ([evalml.objectives.CostBenefitMatrix](#) [method](#)), 368
[objective_function\(\)](#) ([evalml.objectives.ExpVariance](#) [method](#)), 370
[objective_function\(\)](#) ([evalml.objectives.F1](#) [method](#)), 372
[objective_function\(\)](#) ([evalml.objectives.F1Macro](#) [method](#)), 374

`objective_function()`
 (`evalml.objectives.F1Micro` method), 376
`objective_function()`
 (`evalml.objectives.F1Weighted` method), 377
`objective_function()`
 (`evalml.objectives.fraud_cost.FraudCost` method), 272
`objective_function()`
 (`evalml.objectives.FraudCost` method), 379
`objective_function()`
 (`evalml.objectives.lead_scoring.LeadScoring` method), 275
`objective_function()`
 (`evalml.objectives.LeadScoring` method), 382
`objective_function()`
 (`evalml.objectives.LogLossBinary` method), 385
`objective_function()`
 (`evalml.objectives.LogLossMulticlass` method), 386
`objective_function()` (`evalml.objectives.MAE` method), 388
`objective_function()` (`evalml.objectives.MAPE` method), 390
`objective_function()`
 (`evalml.objectives.MaxError` method), 391
`objective_function()`
 (`evalml.objectives.MCCBinary` method), 393
`objective_function()`
 (`evalml.objectives.MCCMulticlass` method), 395
`objective_function()`
 (`evalml.objectives.MeanSquaredLogError` method), 396
`objective_function()`
 (`evalml.objectives.MedianAE` method), 398
`objective_function()` (`evalml.objectives.MSE` method), 400
`objective_function()`
 (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` class method), 277
`objective_function()`
 (`evalml.objectives.MulticlassClassificationObjective` class method), 401
`objective_function()`
 (`evalml.objectives.objective_base.ObjectiveBase` class method), 279
`objective_function()`
 (`evalml.objectives.ObjectiveBase` class method), 403
`objective_function()`
 (`evalml.objectives.Precision` method), 406
`objective_function()`
 (`evalml.objectives.PrecisionMacro` method), 407
`objective_function()`
 (`evalml.objectives.PrecisionMicro` method), 409
`objective_function()`
 (`evalml.objectives.PrecisionWeighted` method), 411
`objective_function()` (`evalml.objectives.R2` method), 412
`objective_function()` (`evalml.objectives.Recall` method), 414
`objective_function()`
 (`evalml.objectives.RecallMacro` method), 416
`objective_function()`
 (`evalml.objectives.RecallMicro` method), 417
`objective_function()`
 (`evalml.objectives.RecallWeighted` method), 419
`objective_function()`
 (`evalml.objectives.regression_objective.RegressionObjective` class method), 282
`objective_function()`
 (`evalml.objectives.RegressionObjective` class method), 421
`objective_function()`
 (`evalml.objectives.RootMeanSquaredError` method), 422
`objective_function()`
 (`evalml.objectives.RootMeanSquaredLogError` method), 424
`objective_function()`
 (`evalml.objectives.sensitivity_low_alert.SensitivityLowAlert` method), 284
`objective_function()`
 (`evalml.objectives.SensitivityLowAlert` method), 426
`objective_function()`
 (`evalml.objectives.standard_metrics.AccuracyBinary` method), 287
`objective_function()`
 (`evalml.objectives.standard_metrics.AccuracyMulticlass` method), 289
`objective_function()`
 (`evalml.objectives.standard_metrics.AUC` method), 291
`objective_function()`
 (`evalml.objectives.standard_metrics.AUCMacro` method), 293
`objective_function()`

<code>(evalml.objectives.standard_metrics.AUCMicro method), 295</code>	<code>(evalml.objectives.standard_metrics.MSE method), 326</code>
<code>objective_function() (evalml.objectives.standard_metrics.AUCWeighted method), 296</code>	<code>objective_function() (evalml.objectives.standard_metrics.Precision method), 327</code>
<code>objective_function() (evalml.objectives.standard_metrics.BalancedAccuracyBinary method), 298</code>	<code>objective_function() (evalml.objectives.standard_metrics.PrecisionMacro method), 329</code>
<code>objective_function() (evalml.objectives.standard_metrics.BalancedAccuracyMulticlass method), 300</code>	<code>objective_function() (evalml.objectives.standard_metrics.PrecisionMicro method), 331</code>
<code>objective_function() (evalml.objectives.standard_metrics.ExpVariance method), 302</code>	<code>objective_function() (evalml.objectives.standard_metrics.PrecisionWeighted method), 332</code>
<code>objective_function() (evalml.objectives.standard_metrics.F1 method), 304</code>	<code>objective_function() (evalml.objectives.standard_metrics.R2 method), 334</code>
<code>objective_function() (evalml.objectives.standard_metrics.F1Macro method), 305</code>	<code>objective_function() (evalml.objectives.standard_metrics.Recall method), 336</code>
<code>objective_function() (evalml.objectives.standard_metrics.F1Micro method), 307</code>	<code>objective_function() (evalml.objectives.standard_metrics.RecallMacro method), 338</code>
<code>objective_function() (evalml.objectives.standard_metrics.F1Weighted method), 309</code>	<code>objective_function() (evalml.objectives.standard_metrics.RecallMicro method), 339</code>
<code>objective_function() (evalml.objectives.standard_metrics.LogLossBinary method), 310</code>	<code>objective_function() (evalml.objectives.standard_metrics.RecallWeighted method), 341</code>
<code>objective_function() (evalml.objectives.standard_metrics.LogLossMulticlass method), 312</code>	<code>objective_function() (evalml.objectives.standard_metrics.RootMeanSquaredError method), 342</code>
<code>objective_function() (evalml.objectives.standard_metrics.MAE method), 314</code>	<code>objective_function() (evalml.objectives.standard_metrics.RootMeanSquaredLogError method), 344</code>
<code>objective_function() (evalml.objectives.standard_metrics.MAPE method), 315</code>	<code>objective_function() (evalml.objectives.time_series_regression_objective.TimeSeriesRe class method), 346</code>
<code>objective_function() (evalml.objectives.standard_metrics.MaxError method), 317</code>	<code>ObjectiveBase (class in evalml.objectives), 402</code>
<code>objective_function() (evalml.objectives.standard_metrics.MCCBinary method), 319</code>	<code>ObjectiveBase (class in evalml.objectives.objective_base), 278</code>
<code>objective_function() (evalml.objectives.standard_metrics.MCCMulticlass method), 321</code>	<code>ObjectiveCreationError, 230</code>
<code>objective_function() (evalml.objectives.standard_metrics.MCCMulticlass method), 321</code>	<code>ObjectiveNotFoundError, 230</code>
<code>objective_function() (evalml.objectives.standard_metrics.MeanSquaredLogError method), 322</code>	<code>OneHotEncoder (class in evalml.pipelines), 1097</code>
<code>objective_function() (evalml.objectives.standard_metrics.MedianAE method), 324</code>	<code>OneHotEncoder (class in evalml.pipelines.components), 925</code>
<code>objective_function()</code>	<code>OneHotEncoder (class in evalml.pipelines.components.transformers), 807</code>
	<code>OneHotEncoder (class in evalml.pipelines.components.transformers.encoders), 660</code>
	<code>OneHotEncoder (class in evalml.pipelines.components.transformers.encoders.onehot_encod</code>

[653](#)
 OneHotEncoderMeta (class in [evalml.pipelines.components.transformers.encoders.one_hot_encoder](#)), 288
[656](#)
 optimize_threshold() (evalml.objectives.AccuracyBinary method), 291
[351](#)
 optimize_threshold() (evalml.objectives.AUC method), 355
 optimize_threshold() (evalml.objectives.BalancedAccuracyBinary method), 362
 optimize_threshold() (evalml.objectives.binary_classification_objective.BinaryClassificationObjective method), 267
 optimize_threshold() (evalml.objectives.BinaryClassificationObjective method), 366
 optimize_threshold() (evalml.objectives.cost_benefit_matrix.CostBenefitMatrix method), 270
 optimize_threshold() (evalml.objectives.CostBenefitMatrix method), 369
 optimize_threshold() (evalml.objectives.F1 method), 372
 optimize_threshold() (evalml.objectives.fraud_cost.FraudCost method), 272
 optimize_threshold() (evalml.objectives.FraudCost method), 380
 optimize_threshold() (evalml.objectives.lead_scoring.LeadScoring method), 275
 optimize_threshold() (evalml.objectives.LeadScoring method), 383
 optimize_threshold() (evalml.objectives.LogLossBinary method), 385
 optimize_threshold() (evalml.objectives.MCCBinary method), 393
 optimize_threshold() (evalml.objectives.Precision method), 406
 optimize_threshold() (evalml.objectives.Recall method), 414
 optimize_threshold() (evalml.objectives.sensitivity_low_alert.SensitivityLowAlert method), 284
 optimize_threshold() (evalml.objectives.SensitivityLowAlert method), 426
 optimize_threshold() (evalml.objectives.standard_metrics.AccuracyBinary method), 288
 optimize_threshold() (evalml.objectives.standard_metrics.AUC method), 291
 optimize_threshold() (evalml.objectives.standard_metrics.BalancedAccuracyBinary method), 298
 optimize_threshold() (evalml.objectives.standard_metrics.F1 method), 304
 optimize_threshold() (evalml.objectives.standard_metrics.LogLossBinary method), 311
 optimize_threshold() (evalml.objectives.standard_metrics.MCCBinary method), 319
 optimize_threshold() (evalml.objectives.standard_metrics.Precision method), 328
 optimize_threshold() (evalml.objectives.standard_metrics.Recall method), 336
 optimize_threshold() (evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline method), 990
 optimize_threshold() (evalml.pipelines.binary_classification_pipeline_mixin.BinaryClassificationPipelineMixin method), 991
 optimize_threshold() (evalml.pipelines.BinaryClassificationPipeline method), 1041
 optimize_threshold() (evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method), 1018
 optimize_threshold() (evalml.pipelines.TimeSeriesBinaryClassificationPipeline method), 1139
 OutliersDataCheck (class in [evalml.data_checks](#)), 221
 OutliersDataCheck (class in [evalml.data_checks.outliers_data_check](#)), 202
P
 pad_with_nans() (in module [evalml.utils](#)), 1197
 pad_with_nans() (in module [evalml.utils.gen_utils](#)), 1191
 ParameterError, 1182, 1184
 parameters() (evalml.pipelines.ARIMARegressor property), 1037
 parameters() (evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline property), 990

`parameters()` (`evalml.pipelines.BinaryClassificationPipeline` property), 1041
`parameters()` (`evalml.pipelines.CatBoostClassifier` property), 1044
`parameters()` (`evalml.pipelines.CatBoostRegressor` property), 1046
`parameters()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` property), 995
`parameters()` (`evalml.pipelines.ClassificationPipeline` property), 1050
`parameters()` (`evalml.pipelines.components.ARIMARegressor` property), 859
`parameters()` (`evalml.pipelines.components.BaselineClassifier` property), 862
`parameters()` (`evalml.pipelines.components.BaselineRegressor` property), 864
`parameters()` (`evalml.pipelines.components.CatBoostClassifier` property), 867
`parameters()` (`evalml.pipelines.components.CatBoostRegressor` property), 869
`parameters()` (`evalml.pipelines.components.component_base.ComponentBase` property), 850
`parameters()` (`evalml.pipelines.components.ComponentBase` property), 871
`parameters()` (`evalml.pipelines.components.DateTimeFeaturizer` property), 874
`parameters()` (`evalml.pipelines.components.DecisionTreeClassifier` property), 877
`parameters()` (`evalml.pipelines.components.DecisionTreeRegressor` property), 880
`parameters()` (`evalml.pipelines.components.DelayedFeatureTransformer` property), 882
`parameters()` (`evalml.pipelines.components.DFSTransformer` property), 884
`parameters()` (`evalml.pipelines.components.DropColumns` property), 886
`parameters()` (`evalml.pipelines.components.DropNullColumns` property), 888
`parameters()` (`evalml.pipelines.components.ElasticNetClassifier` property), 891
`parameters()` (`evalml.pipelines.components.ElasticNetRegressor` property), 893
`parameters()` (`evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase` property), 430
`parameters()` (`evalml.pipelines.components.ensemble.stacked_ensemble_classifier.StackedEnsembleClassifier` property), 433
`parameters()` (`evalml.pipelines.components.ensemble.stacked_ensemble_regressor.StackedEnsembleRegressor` property), 436
`parameters()` (`evalml.pipelines.components.ensemble.StackedEnsembleBase` property), 439
`parameters()` (`evalml.pipelines.components.ensemble.StackedEnsembleClassifier` property), 441
`parameters()` (`evalml.pipelines.components.ensemble.StackedEnsembleRegressor` property), 444
`parameters()` (`evalml.pipelines.components.Estimator` property), 896
`parameters()` (`evalml.pipelines.components.estimators.ARIMARegressor` property), 582
`parameters()` (`evalml.pipelines.components.estimators.BaselineClassifier` property), 585
`parameters()` (`evalml.pipelines.components.estimators.BaselineRegressor` property), 587
`parameters()` (`evalml.pipelines.components.estimators.CatBoostClassifier` property), 590
`parameters()` (`evalml.pipelines.components.estimators.CatBoostRegressor` property), 592
`parameters()` (`evalml.pipelines.components.estimators.classifiers.baseline.BaselineClassifier` property), 447
`parameters()` (`evalml.pipelines.components.estimators.classifiers.BaselineClassifier` property), 480
`parameters()` (`evalml.pipelines.components.estimators.classifiers.catboost.CatBoostClassifier` property), 450
`parameters()` (`evalml.pipelines.components.estimators.classifiers.CatBoostClassifier` property), 483
`parameters()` (`evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier` property), 453
`parameters()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier` property), 486
`parameters()` (`evalml.pipelines.components.estimators.classifiers.elasticnet.ElasticNetClassifier` property), 456
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier` property), 489
`parameters()` (`evalml.pipelines.components.estimators.classifiers.et_classifier.DecisionTreeClassifier` property), 459
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier` property), 492
`parameters()` (`evalml.pipelines.components.estimators.classifiers.kneighbors.KNeighborsClassifier` property), 462
`parameters()` (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier` property), 495
`parameters()` (`evalml.pipelines.components.estimators.classifiers.lightgbm.LightGBMClassifier` property), 466
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier` property), 498
`parameters()` (`evalml.pipelines.components.estimators.classifiers.logistic.LogisticClassifier` property), 469
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LogisticClassifier` property), 501
`parameters()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier` property), 503
`parameters()` (`evalml.pipelines.components.estimators.classifiers.rf_classifier.RandomForestClassifier` property), 472
`parameters()` (`evalml.pipelines.components.estimators.classifiers.svm.SupportVectorMachineClassifier` property), 475
`parameters()` (`evalml.pipelines.components.estimators.classifiers.SVMClassifier` property), 506
`parameters()` (`evalml.pipelines.components.estimators.classifiers.xgboost.XGBoostClassifier` property), 478

`parameters()` (`evalml.pipelines.components.estimators.classifiers.XGBClassifier`), 509
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier`), 509
`parameters()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier`), 595
`parameters()` (`evalml.pipelines.components.estimators.classifiers.DecisionTreeRegressor`), 598
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier`), 601
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor`), 603
`parameters()` (`evalml.pipelines.components.estimators.classifiers.Estimator`), 606
`parameters()` (`evalml.pipelines.components.estimators.classifiers.estimator.Estimator`), 579
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier`), 609
`parameters()` (`evalml.pipelines.components.estimators.classifiers.ExtraTreesRegressor`), 612
`parameters()` (`evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier`), 614
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LightGBMClassifier`), 617
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LightGBMRegressor`), 620
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LinearRegressor`), 622
`parameters()` (`evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier`), 625
`parameters()` (`evalml.pipelines.components.estimators.classifiers.RandomForestClassifier`), 628
`parameters()` (`evalml.pipelines.components.estimators.classifiers.RandomForestRegressor`), 630
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.arnim()`), 512
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.ARIMARegressor`), 547
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.baseline`), 515
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.BaselineRegressor`), 550
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.catboost`), 518
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.CatBoostRegressor`), 552
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.decisiontree`), 521
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.Desist()`), 555
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.elasticnet`), 524
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.ElasticNetRegressor`), 558
`parameters()` (`evalml.pipelines.components.estimators.classifiers.regressor.et_regressor`), 527

[parameters\(\) \(evalml.pipelines.components.LogisticRegressionClassifier\), 922](#)
[parameters\(\) \(evalml.pipelines.components.LSA\), 924](#)
[parameters\(\) \(evalml.pipelines.components.OneHotEncoder\), 927](#)
[parameters\(\) \(evalml.pipelines.components.PCA\), 929](#)
[parameters\(\) \(evalml.pipelines.components.PerColumnTransformer\), 932](#)
[parameters\(\) \(evalml.pipelines.components.PolynomialDeterminer\), 934](#)
[parameters\(\) \(evalml.pipelines.components.RandomForestClassifier\), 936](#)
[parameters\(\) \(evalml.pipelines.components.RandomForestRegressor\), 939](#)
[parameters\(\) \(evalml.pipelines.components.RFClassifierSelectFromModel\), 941](#)
[parameters\(\) \(evalml.pipelines.components.RFRegressorSelectFromModel\), 944](#)
[parameters\(\) \(evalml.pipelines.components.SelectColumnTransformer\), 946](#)
[parameters\(\) \(evalml.pipelines.components.SimpleImputer\), 948](#)
[parameters\(\) \(evalml.pipelines.components.SMOTENCWrapper\), 951](#)
[parameters\(\) \(evalml.pipelines.components.SMOTENCWrapper\), 953](#)
[parameters\(\) \(evalml.pipelines.components.SMOTESampler\), 955](#)
[parameters\(\) \(evalml.pipelines.components.StackedEnsembleClassifier\), 958](#)
[parameters\(\) \(evalml.pipelines.components.StackedEnsembleRegressor\), 960](#)
[parameters\(\) \(evalml.pipelines.components.StandardScaler\), 962](#)
[parameters\(\) \(evalml.pipelines.components.SVMClassifier\), 964](#)
[parameters\(\) \(evalml.pipelines.components.SVMRegressor\), 967](#)
[parameters\(\) \(evalml.pipelines.components.TargetEncoder\), 969](#)
[parameters\(\) \(evalml.pipelines.components.TargetImputer\), 971](#)
[parameters\(\) \(evalml.pipelines.components.TextFeaturizer\), 973](#)
[parameters\(\) \(evalml.pipelines.components.TimeSeriesBaselineEstimator\), 976](#)
[parameters\(\) \(evalml.pipelines.components.Transformer\), 978](#)
[parameters\(\) \(evalml.pipelines.components.transformers.column_selector.CaliforniaHousingSelector\), 776](#)
[parameters\(\) \(evalml.pipelines.components.transformers.column_selector.CaliforniaHousingSelector\), 778](#)
[parameters\(\) \(evalml.pipelines.components.transformers.column_selector.CaliforniaHousingSelector\), 780](#)
[parameters\(\) \(evalml.pipelines.components.transformers.DateTimeFeaturizer\), 789](#)
[parameters\(\) \(evalml.pipelines.components.transformers.DelayedFeaturizer\), 791](#)
[parameters\(\) \(evalml.pipelines.components.transformers.DFSTransformer\), 793](#)
[parameters\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityFeaturizer\), 645](#)
[parameters\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityFeaturizer\), 650](#)
[parameters\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityFeaturizer\), 652](#)
[parameters\(\) \(evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimensionalityFeaturizer\), 647](#)
[parameters\(\) \(evalml.pipelines.components.transformers.DropColumnTransformer\), 795](#)
[parameters\(\) \(evalml.pipelines.components.transformers.DropNullColumnTransformer\), 797](#)
[parameters\(\) \(evalml.pipelines.components.transformers.encoders.one_hot.OneHotEncoder\), 655](#)
[parameters\(\) \(evalml.pipelines.components.transformers.encoders.OneHotEncoder\), 662](#)
[parameters\(\) \(evalml.pipelines.components.transformers.encoders.target.TargetEncoder\), 659](#)
[parameters\(\) \(evalml.pipelines.components.transformers.encoders.TargetEncoder\), 665](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 668](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 677](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 671](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 674](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 680](#)
[parameters\(\) \(evalml.pipelines.components.transformers.feature_selector.FeatureSelector\), 683](#)
[parameters\(\) \(evalml.pipelines.components.transformers.FeatureSelector\), 800](#)
[parameters\(\) \(evalml.pipelines.components.transformers.Imputer\), 802](#)
[parameters\(\) \(evalml.pipelines.components.transformers.imputers.Imputer\), 697](#)
[parameters\(\) \(evalml.pipelines.components.transformers.imputers.imputer.Imputer\), 686](#)
[parameters\(\) \(evalml.pipelines.components.transformers.imputers.per_column.PerColumnImputer\), 688](#)
[parameters\(\) \(evalml.pipelines.components.transformers.imputers.PerColumnImputer\), 699](#)
[parameters\(\) \(evalml.pipelines.components.transformers.imputers.SimpleImputer\), 691](#)

`parameters()` (`evalml.pipelines.components.XGBoostRegressor` property), 985
`parameters()` (`evalml.pipelines.DecisionTreeClassifier` property), 1056
`parameters()` (`evalml.pipelines.DecisionTreeRegressor` property), 1059
`parameters()` (`evalml.pipelines.DelayedFeatureTransformer` property), 1061
`parameters()` (`evalml.pipelines.DFSTransformer` property), 1064
`parameters()` (`evalml.pipelines.ElasticNetClassifier` property), 1066
`parameters()` (`evalml.pipelines.ElasticNetRegressor` property), 1069
`parameters()` (`evalml.pipelines.Estimator` property), 1071
`parameters()` (`evalml.pipelines.ExtraTreesClassifier` property), 1074
`parameters()` (`evalml.pipelines.ExtraTreesRegressor` property), 1077
`parameters()` (`evalml.pipelines.FeatureSelector` property), 1079
`parameters()` (`evalml.pipelines.KNeighborsClassifier` property), 1082
`parameters()` (`evalml.pipelines.LightGBMClassifier` property), 1085
`parameters()` (`evalml.pipelines.LightGBMRegressor` property), 1087
`parameters()` (`evalml.pipelines.LinearRegressor` property), 1090
`parameters()` (`evalml.pipelines.LogisticRegressionClassifier` property), 1092
`parameters()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` property), 1003
`parameters()` (`evalml.pipelines.MulticlassClassificationPipeline` property), 1096
`parameters()` (`evalml.pipelines.OneHotEncoder` property), 1099
`parameters()` (`evalml.pipelines.PerColumnImputer` property), 1102
`parameters()` (`evalml.pipelines.pipeline_base.PipelineBase` property), 1008
`parameters()` (`evalml.pipelines.PipelineBase` property), 1105
`parameters()` (`evalml.pipelines.RandomForestClassifier` property), 1108
`parameters()` (`evalml.pipelines.RandomForestRegressor` property), 1110
`parameters()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` property), 1013
`parameters()` (`evalml.pipelines.RegressionPipeline` property), 1114
`parameters()` (`evalml.pipelines.RFClassifierSelectFromModel` property), 1116
`parameters()` (`evalml.pipelines.RFRegressorSelectFromModel` property), 1119
`parameters()` (`evalml.pipelines.SimpleImputer` property), 1121
`parameters()` (`evalml.pipelines.StackedEnsembleClassifier` property), 1124
`parameters()` (`evalml.pipelines.StackedEnsembleRegressor` property), 1126
`parameters()` (`evalml.pipelines.StandardScaler` property), 1129
`parameters()` (`evalml.pipelines.SVMClassifier` property), 1131
`parameters()` (`evalml.pipelines.SVMRegressor` property), 1133
`parameters()` (`evalml.pipelines.TargetEncoder` property), 1136
`parameters()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline` property), 1018
`parameters()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline` property), 1022
`parameters()` (`evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline` property), 1027
`parameters()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` property), 1031
`parameters()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` property), 1140
`parameters()` (`evalml.pipelines.TimeSeriesClassificationPipeline` property), 1144
`parameters()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` property), 1148
`parameters()` (`evalml.pipelines.TimeSeriesRegressionPipeline` property), 1152
`parameters()` (`evalml.pipelines.TimeSeriesRegressionPipeline` property), 1155
`parameters()` (`evalml.pipelines.XGBoostClassifier` property), 1157
`parameters()` (`evalml.pipelines.XGBoostRegressor` property), 1159
`partial_dependence()` (in module `evalml.model_understanding`), 262
`partial_dependence()` (in module `evalml.model_understanding.graphs`), 249
PCA (class in `evalml.pipelines.components`), 928
PCA (class in `evalml.pipelines.components.transformers`), 810
PCA (class in `evalml.pipelines.components.transformers.dimensionality_reduction`), 650
PCA (class in `evalml.pipelines.components.transformers.dimensionality_reduction`), 650
PerColumnImputer (class in `evalml.pipelines`), 1100
PerColumnImputer (class in `evalml.pipelines.components`), 930
PerColumnImputer (class in `evalml.pipelines.components.transformers`), 930

[812](#)
 PerColumnImputer (class in [evalml.pipelines.components.transformers.imputer](#)), [814](#)
[697](#)
 PerColumnImputer (class in [evalml.pipelines.components.transformers.imputer](#)), [737](#)
[687](#)
 perfect_score() (evalml.objectives.binary_classification_objective.BinaryClassificationObjective property), [267](#)
 perfect_score() (evalml.objectives.BinaryClassificationObjective method), [352](#)
 perfect_score() (evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective property), [277](#)
 perfect_score() (evalml.objectives.MulticlassClassificationObjective method), [355](#)
 perfect_score() (evalml.objectives.objective_base.ObjectiveBase method), [357](#)
 perfect_score() (evalml.objectives.ObjectiveBase property), [280](#)
 perfect_score() (evalml.objectives.ObjectiveBase method), [359](#)
 perfect_score() (evalml.objectives.regression_objective.RegressionObjective property), [282](#)
 perfect_score() (evalml.objectives.RegressionObjective method), [362](#)
 perfect_score() (evalml.objectives.RegressionObjective property), [421](#)
 perfect_score() (evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective property), [346](#)
 pipeline_number() (evalml.automl.automl_algorithm.automl_algorithm.AutoMLAlgorithm property), [145](#)
 pipeline_number() (evalml.automl.automl_algorithm.AutoMLAlgorithm method), [270](#)
 pipeline_number() (evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm property), [147](#)
 pipeline_number() (evalml.automl.automl_algorithm.iterative_algorithm.IterativeAlgorithm method), [373](#)
 PipelineBase (class in [evalml.pipelines](#)), [1102](#)
 PipelineBase (class in [evalml.pipelines.pipeline_base](#)), [1005](#)
 PipelineBaseMeta (class in [evalml.pipelines.pipeline_meta](#)), [1009](#)
 PipelineNotFoundError, [230](#)
 PipelineNotYetFittedError, [230](#), [231](#)
 PipelineScoreError, [230](#), [231](#)
 PipelineSearchPlots (class in [evalml.automl.pipeline_search_plots](#)), [172](#)
 plot() (evalml.automl.automl_search.AutoMLSearch property), [170](#)
 plot() (evalml.automl.AutoMLSearch property), [180](#)
 plot() (evalml.AutoMLSearch property), [1202](#)
 PolynomialDetrender (class in [evalml.pipelines.components](#)), [932](#)
 PolynomialDetrender (class in [evalml.pipelines.components.transformers](#)), [814](#)
 PolynomialDetrender (class in [evalml.pipelines.components.transformers.preprocessing](#)), [737](#)
 PolynomialDetrender (class in [evalml.pipelines.components.transformers.preprocessing.polynomial](#)), [718](#)
 positive_only() (evalml.objectives.AccuracyBinary method), [352](#)
 positive_only() (evalml.objectives.AccuracyMulticlass method), [355](#)
 positive_only() (evalml.objectives.AUC method), [355](#)
 positive_only() (evalml.objectives.AUCMacro method), [359](#)
 positive_only() (evalml.objectives.AUCMicro method), [359](#)
 positive_only() (evalml.objectives.AUCWeighted method), [360](#)
 positive_only() (evalml.objectives.BalancedAccuracyBinary method), [362](#)
 positive_only() (evalml.objectives.BalancedAccuracyMulticlass method), [369](#)
 positive_only() (evalml.objectives.binary_classification_objective.BinaryClassificationObjective method), [267](#)
 positive_only() (evalml.objectives.cost_benefit_matrix.CostBenefitMatrix method), [270](#)
 positive_only() (evalml.objectives.CostBenefitMatrix method), [369](#)
 positive_only() (evalml.objectives.ExpVariance method), [370](#)
 positive_only() (evalml.objectives.F1 method), [373](#)
 positive_only() (evalml.objectives.F1Macro method), [374](#)
 positive_only() (evalml.objectives.F1Micro method), [376](#)
 positive_only() (evalml.objectives.F1Weighted method), [377](#)
 positive_only() (evalml.objectives.fraud_cost.FraudCost method), [273](#)
 positive_only() (evalml.objectives.FraudCost method), [380](#)
 positive_only() (evalml.objectives.lead_scoring.LeadScoring method), [275](#)
 positive_only() (evalml.objectives.LeadScoring method), [383](#)
 positive_only() (evalml.objectives.LogLossBinary method), [385](#)
 positive_only() (evalml.objectives.LogLossMulticlass method), [387](#)

`positive_only()` (*evalml.objectives.MAE method*), 388
`positive_only()` (*evalml.objectives.MAPE method*), 390
`positive_only()` (*evalml.objectives.MaxError method*), 391
`positive_only()` (*evalml.objectives.MCCBinary method*), 393
`positive_only()` (*evalml.objectives.MCCMulticlass method*), 395
`positive_only()` (*evalml.objectives.MeanSquaredLogError method*), 397
`positive_only()` (*evalml.objectives.MedianAE method*), 398
`positive_only()` (*evalml.objectives.MSE method*), 400
`positive_only()` (*evalml.objectives.multiclass_classification_objective.MulticlassLogLossObjective method*), 277
`positive_only()` (*evalml.objectives.MulticlassClassificationObjective method*), 402
`positive_only()` (*evalml.objectives.objective_base.ObjectiveBase method*), 280
`positive_only()` (*evalml.objectives.ObjectiveBase method*), 404
`positive_only()` (*evalml.objectives.Precision method*), 406
`positive_only()` (*evalml.objectives.PrecisionMacro method*), 408
`positive_only()` (*evalml.objectives.PrecisionMicro method*), 409
`positive_only()` (*evalml.objectives.PrecisionWeighted method*), 411
`positive_only()` (*evalml.objectives.R2 method*), 412
`positive_only()` (*evalml.objectives.Recall method*), 414
`positive_only()` (*evalml.objectives.RecallMacro method*), 416
`positive_only()` (*evalml.objectives.RecallMicro method*), 418
`positive_only()` (*evalml.objectives.RecallWeighted method*), 419
`positive_only()` (*evalml.objectives.regression_objective.RegressionObjective method*), 282
`positive_only()` (*evalml.objectives.RegressionObjective method*), 421
`positive_only()` (*evalml.objectives.RootMeanSquaredError method*), 423
`positive_only()` (*evalml.objectives.RootMeanSquaredLogError method*), 424
`positive_only()` (*evalml.objectives.sensitivity_low_alert.SensitivityLowAlert method*), 284
`positive_only()` (*evalml.objectives.SensitivityLowAlert method*), 426
`positive_only()` (*evalml.objectives.standard_metrics.AccuracyBinary method*), 288
`positive_only()` (*evalml.objectives.standard_metrics.AccuracyMulticlass method*), 289
`positive_only()` (*evalml.objectives.standard_metrics.AUC method*), 292
`positive_only()` (*evalml.objectives.standard_metrics.AUCMacro method*), 293
`positive_only()` (*evalml.objectives.standard_metrics.AUCMicro method*), 295
`positive_only()` (*evalml.objectives.standard_metrics.AUCWeighted method*), 296
`positive_only()` (*evalml.objectives.standard_metrics.BalancedAccuracyBinary method*), 299
`positive_only()` (*evalml.objectives.standard_metrics.BalancedAccuracyMulticlass method*), 300
`positive_only()` (*evalml.objectives.standard_metrics.ExpVariance method*), 302
`positive_only()` (*evalml.objectives.standard_metrics.F1 method*), 304
`positive_only()` (*evalml.objectives.standard_metrics.F1Macro method*), 306
`positive_only()` (*evalml.objectives.standard_metrics.F1Micro method*), 307
`positive_only()` (*evalml.objectives.standard_metrics.F1Weighted method*), 309
`positive_only()` (*evalml.objectives.standard_metrics.LogLossBinary method*), 311
`positive_only()` (*evalml.objectives.standard_metrics.LogLossMulticlass method*), 312
`positive_only()` (*evalml.objectives.standard_metrics.MAE method*), 314
`positive_only()` (*evalml.objectives.standard_metrics.MAPE method*), 316
`positive_only()` (*evalml.objectives.standard_metrics.MaxError method*), 317
`positive_only()` (*evalml.objectives.standard_metrics.MCCBinary method*), 319
`positive_only()` (*evalml.objectives.standard_metrics.MCCMulticlass method*), 321
`positive_only()` (*evalml.objectives.standard_metrics.MeanSquaredLogError method*), 323
`positive_only()` (*evalml.objectives.standard_metrics.MedianAE method*), 324
`positive_only()` (*evalml.objectives.standard_metrics.MSE method*), 326
`positive_only()` (*evalml.objectives.standard_metrics.Precision method*), 328
`positive_only()` (*evalml.objectives.standard_metrics.PrecisionMacro method*), 329
`positive_only()` (*evalml.objectives.standard_metrics.PrecisionMicro method*), 331
`positive_only()` (*evalml.objectives.standard_metrics.PrecisionWeighted method*), 333

`positive_only()` (*evalml.objectives.standard_metrics.R2* method), 862
 method), 334
`positive_only()` (*evalml.objectives.standard_metrics.Recall* method), 864
 method), 336
`positive_only()` (*evalml.objectives.standard_metrics.RecallMacro* method), 867
 method), 338
`positive_only()` (*evalml.objectives.standard_metrics.RecallMicro* method), 869
 method), 339
`positive_only()` (*evalml.objectives.standard_metrics.RecallWeighted* method), 877
 method), 341
`positive_only()` (*evalml.objectives.standard_metrics.RootMeanSquaredError* method), 880
 method), 343
`positive_only()` (*evalml.objectives.standard_metrics.RootMeanSquaredLogError* method), 889
 method), 344
`positive_only()` (*evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective* method), 899
 method), 346
Precision (class in *evalml.objectives*), 404
Precision (class in *evalml.objectives.standard_metrics*), 326
`precision_recall_curve()` (in module *evalml.model_understanding*), 263
`precision_recall_curve()` (in module *evalml.model_understanding.graphs*), 250
PrecisionMacro (class in *evalml.objectives*), 407
PrecisionMacro (class in *evalml.objectives.standard_metrics*), 328
PrecisionMicro (class in *evalml.objectives*), 408
PrecisionMicro (class in *evalml.objectives.standard_metrics*), 330
PrecisionWeighted (class in *evalml.objectives*), 410
PrecisionWeighted (class in *evalml.objectives.standard_metrics*), 332
`predict()` (*evalml.pipelines.ARIMAREgressor* method), 1037
`predict()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 990
`predict()` (*evalml.pipelines.BinaryClassificationPipeline* method), 1041
`predict()` (*evalml.pipelines.CatBoostClassifier* method), 1044
`predict()` (*evalml.pipelines.CatBoostRegressor* method), 1046
`predict()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 995
`predict()` (*evalml.pipelines.ClassificationPipeline* method), 1050
`predict()` (*evalml.pipelines.component_graph.ComponentGraph* method), 1000
`predict()` (*evalml.pipelines.ComponentGraph* method), 1054
`predict()` (*evalml.pipelines.components.ARIMAREgressor* method), 859
`predict()` (*evalml.pipelines.components.BaselineClassifier* method), 862
`predict()` (*evalml.pipelines.components.BaselineRegressor* method), 864
`predict()` (*evalml.pipelines.components.CatBoostClassifier* method), 867
`predict()` (*evalml.pipelines.components.CatBoostRegressor* method), 869
`predict()` (*evalml.pipelines.components.DecisionTreeClassifier* method), 877
`predict()` (*evalml.pipelines.components.DecisionTreeRegressor* method), 880
`predict()` (*evalml.pipelines.components.ElasticNetClassifier* method), 889
`predict()` (*evalml.pipelines.components.ElasticNetRegressor* method), 899
`predict()` (*evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase* method), 430
`predict()` (*evalml.pipelines.components.ensemble.stacked_ensemble_classifier.StackedEnsembleClassifier* method), 433
`predict()` (*evalml.pipelines.components.ensemble.stacked_ensemble_regressor.StackedEnsembleRegressor* method), 436
`predict()` (*evalml.pipelines.components.ensemble.StackedEnsembleBase* method), 439
`predict()` (*evalml.pipelines.components.ensemble.StackedEnsembleClassifier* method), 441
`predict()` (*evalml.pipelines.components.ensemble.StackedEnsembleRegressor* method), 444
`predict()` (*evalml.pipelines.components.Estimator* method), 896
`predict()` (*evalml.pipelines.components.estimators.ARIMAREgressor* method), 582
`predict()` (*evalml.pipelines.components.estimators.BaselineClassifier* method), 585
`predict()` (*evalml.pipelines.components.estimators.BaselineRegressor* method), 587
`predict()` (*evalml.pipelines.components.estimators.CatBoostClassifier* method), 590
`predict()` (*evalml.pipelines.components.estimators.CatBoostRegressor* method), 592
`predict()` (*evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier* method), 447
`predict()` (*evalml.pipelines.components.estimators.classifiers.BaselineRegressor* method), 480
`predict()` (*evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier* method), 450
`predict()` (*evalml.pipelines.components.estimators.classifiers.CatBoostRegressor* method), 483
`predict()` (*evalml.pipelines.components.estimators.classifiers.decision_tree.DecisionTreeClassifier* method), 453
`predict()` (*evalml.pipelines.components.estimators.classifiers.DecisionTreeRegressor* method), 486
`predict()` (*evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.ElasticNetClassifier* method), 456
`predict()` (*evalml.pipelines.components.estimators.classifiers.ElasticNetRegressor* method), 489

method), 489

predict () (evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier), 459

predict () (evalml.pipelines.components.estimators.classifiers.ExtraTreeClassifier), 492

predict () (evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier), 463

predict () (evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier), 495

predict () (evalml.pipelines.components.estimators.classifiers.LightGBMClassifier), 466

predict () (evalml.pipelines.components.estimators.classifiers.LightGBMClassifier), 498

predict () (evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier), 469

predict () (evalml.pipelines.components.estimators.classifiers.LogisticRegressionClassifier), 501

predict () (evalml.pipelines.components.estimators.classifiers.RandomForestClassifier), 503

predict () (evalml.pipelines.components.estimators.classifiers.SVCClassifier), 472

predict () (evalml.pipelines.components.estimators.classifiers.SVCClassifier), 475

predict () (evalml.pipelines.components.estimators.classifiers.SVCClassifier), 506

predict () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier), 478

predict () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier), 509

predict () (evalml.pipelines.components.estimators.DecisionTreeClassifier), 595

predict () (evalml.pipelines.components.estimators.DecisionTreeRegressor), 598

predict () (evalml.pipelines.components.estimators.ElasticNetClassifier), 601

predict () (evalml.pipelines.components.estimators.ElasticNetRegressor), 604

predict () (evalml.pipelines.components.estimators.Estimator), 606

predict () (evalml.pipelines.components.estimators.estimator.Estimator), 579

predict () (evalml.pipelines.components.estimators.ExtraTreeClassifier), 609

predict () (evalml.pipelines.components.estimators.ExtraTreeRegressor), 612

predict () (evalml.pipelines.components.estimators.KNeighborsClassifier), 614

predict () (evalml.pipelines.components.estimators.LightGBMClassifier), 617

predict () (evalml.pipelines.components.estimators.LightGBMRegressor), 620

predict () (evalml.pipelines.components.estimators.LinearRegressor), 623

predict () (evalml.pipelines.components.estimators.LogisticRegressionClassifier), 625

predict () (evalml.pipelines.components.estimators.LogisticRegressionClassifier), 628

predict () (evalml.pipelines.components.estimators.RandomForestClassifier), 630

predict () (evalml.pipelines.components.estimators.regressors.arima_regressor), 512

predict () (evalml.pipelines.components.estimators.regressors.ARIMAResidualsRegressor), 547

predict () (evalml.pipelines.components.estimators.regressors.baseline_regressor), 515

predict () (evalml.pipelines.components.estimators.regressors.BaselineRegressor), 550

predict () (evalml.pipelines.components.estimators.regressors.catboost_regressor), 518

predict () (evalml.pipelines.components.estimators.regressors.CatBoostRegressor), 552

predict () (evalml.pipelines.components.estimators.regressors.decision_tree_regressor), 521

predict () (evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor), 555

predict () (evalml.pipelines.components.estimators.regressors.elasticnet_regressor), 524

predict () (evalml.pipelines.components.estimators.regressors.ElasticNetRegressor), 558

predict () (evalml.pipelines.components.estimators.regressors.et_regressor), 527

predict () (evalml.pipelines.components.estimators.regressors.ExtraTreeRegressor), 561

predict () (evalml.pipelines.components.estimators.regressors.lightgbm_regressor), 530

predict () (evalml.pipelines.components.estimators.regressors.LightGBMRegressor), 564

predict () (evalml.pipelines.components.estimators.regressors.linear_regressor), 533

predict () (evalml.pipelines.components.estimators.regressors.LinearRegressor), 566

predict () (evalml.pipelines.components.estimators.regressors.RandomForestRegressor), 568

predict () (evalml.pipelines.components.estimators.regressors.rf_regressor), 536

predict () (evalml.pipelines.components.estimators.regressors.svm_regressor), 538

predict () (evalml.pipelines.components.estimators.regressors.SVMRegressor), 571

predict () (evalml.pipelines.components.estimators.regressors.time_series_regressor), 541

predict () (evalml.pipelines.components.estimators.regressors.TimeSeriesRegressor), 573

predict () (evalml.pipelines.components.estimators.regressors.xgboost_regressor), 544

predict () (evalml.pipelines.components.estimators.regressors.XGBoostRegressor), 576

predict () (evalml.pipelines.components.estimators.SVMClassifier), 576

`predict()` (`evalml.pipelines.XGBoostClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.BaselineRe`
`method`), 1157 `method`), 587
`predict()` (`evalml.pipelines.XGBoostRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.CatBoostC`
`method`), 1159 `method`), 590
`predict_proba()` (`evalml.pipelines.ARIMARegressor` `predict_proba()` (`evalml.pipelines.components.estimators.CatBoostR`
`method`), 1037 `method`), 592
`predict_proba()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 990 `method`), 447
`predict_proba()` (`evalml.pipelines.BinaryClassificationPipeline` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 1041 `method`), 481
`predict_proba()` (`evalml.pipelines.CatBoostClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.c`
`method`), 1044 `method`), 450
`predict_proba()` (`evalml.pipelines.CatBoostRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.c`
`method`), 1046 `method`), 483
`predict_proba()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.c`
`method`), 995 `method`), 453
`predict_proba()` (`evalml.pipelines.ClassificationPipeline` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 1050 `method`), 486
`predict_proba()` (`evalml.pipelines.components.ARIMARegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.e`
`method`), 859 `method`), 456
`predict_proba()` (`evalml.pipelines.components.BaselineClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 862 `method`), 489
`predict_proba()` (`evalml.pipelines.components.BaselineRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.e`
`method`), 864 `method`), 460
`predict_proba()` (`evalml.pipelines.components.CatBoostClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 867 `method`), 492
`predict_proba()` (`evalml.pipelines.components.CatBoostRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.k`
`method`), 869 `method`), 463
`predict_proba()` (`evalml.pipelines.components.DecisionTreeClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 877 `method`), 495
`predict_proba()` (`evalml.pipelines.components.DecisionTreeRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 880 `method`), 466
`predict_proba()` (`evalml.pipelines.components.ElasticNetClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 891 `method`), 498
`predict_proba()` (`evalml.pipelines.components.ElasticNetRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 894 `method`), 469
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.BaseStackedEnsembleBase` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 430 `method`), 501
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.ClassifiedStackedEnsembleClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.l`
`method`), 433 `method`), 504
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.RegressedStackedEnsembleRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.n`
`method`), 436 `method`), 472
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.BaseStackedEnsembleBase` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.s`
`method`), 439 `method`), 475
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.ClassifiedStackedEnsembleClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.s`
`method`), 441 `method`), 506
`predict_proba()` (`evalml.pipelines.components.ensemble_stacked_ensemble.RegressedStackedEnsembleRegressor` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.x`
`method`), 444 `method`), 478
`predict_proba()` (`evalml.pipelines.components.Estimator` `predict_proba()` (`evalml.pipelines.components.estimators.classifiers.x`
`method`), 896 `method`), 509
`predict_proba()` (`evalml.pipelines.components.estimators.ARIMARegressor` `predict_proba()` (`evalml.pipelines.components.estimators.DecisionTr`
`method`), 582 `method`), 595
`predict_proba()` (`evalml.pipelines.components.estimators.BaselineClassifier` `predict_proba()` (`evalml.pipelines.components.estimators.DecisionTr`
`method`), 585 `method`), 598

`predict_proba()` (`evalml.pipelines.components.estimators.elasticnet_classifier`), 601
`predict_proba()` (`evalml.pipelines.components.estimators.elasticnet_regressor`), 604
`predict_proba()` (`evalml.pipelines.components.estimators.estimator`), 606
`predict_proba()` (`evalml.pipelines.components.estimators.estimator`), 579
`predict_proba()` (`evalml.pipelines.components.estimators.extra_trees_classifier`), 609
`predict_proba()` (`evalml.pipelines.components.estimators.extra_trees_regressor`), 612
`predict_proba()` (`evalml.pipelines.components.estimators.kneighbors_classifier`), 614
`predict_proba()` (`evalml.pipelines.components.estimators.lightgbm_classifier`), 617
`predict_proba()` (`evalml.pipelines.components.estimators.lightgbm_regressor`), 620
`predict_proba()` (`evalml.pipelines.components.estimators.linear_regressor`), 623
`predict_proba()` (`evalml.pipelines.components.estimators.logistic_regression_classifier`), 625
`predict_proba()` (`evalml.pipelines.components.estimators.random_forest_classifier`), 628
`predict_proba()` (`evalml.pipelines.components.estimators.random_forest_regressor`), 630
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 512
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 547
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 515
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 550
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 518
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 552
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 521
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 555
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 524
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 558
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 527
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 561
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 530
`predict_proba()` (`evalml.pipelines.components.estimators.regressor`), 564

[predict_proba\(\) \(evalml.pipelines.components.SVMRegressor method\), 967](#)
[predict_proba\(\) \(evalml.pipelines.components.TimeSeriesBaselineEstimator method\), 976](#)
[predict_proba\(\) \(evalml.pipelines.components.utils.WrappedSKClassifier method\), 853](#)
[predict_proba\(\) \(evalml.pipelines.components.XGBoostClassifier method\), 983](#)
[predict_proba\(\) \(evalml.pipelines.components.XGBoostRegressor method\), 986](#)
[predict_proba\(\) \(evalml.pipelines.DecisionTreeClassifier method\), 1056](#)
[predict_proba\(\) \(evalml.pipelines.DecisionTreeRegressor method\), 1059](#)
[predict_proba\(\) \(evalml.pipelines.ElasticNetClassifier method\), 1067](#)
[predict_proba\(\) \(evalml.pipelines.ElasticNetRegressor method\), 1069](#)
[predict_proba\(\) \(evalml.pipelines.Estimator method\), 1071](#)
[predict_proba\(\) \(evalml.pipelines.ExtraTreesClassifier method\), 1074](#)
[predict_proba\(\) \(evalml.pipelines.ExtraTreesRegressor method\), 1077](#)
[predict_proba\(\) \(evalml.pipelines.KNeighborsClassifier method\), 1082](#)
[predict_proba\(\) \(evalml.pipelines.LightGBMClassifier method\), 1085](#)
[predict_proba\(\) \(evalml.pipelines.LightGBMRegressor method\), 1088](#)
[predict_proba\(\) \(evalml.pipelines.LinearRegressor method\), 1090](#)
[predict_proba\(\) \(evalml.pipelines.LogisticRegressionClassifier method\), 1092](#)
[predict_proba\(\) \(evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method\), 1004](#)
[predict_proba\(\) \(evalml.pipelines.MulticlassClassificationPipeline method\), 1096](#)
[predict_proba\(\) \(evalml.pipelines.RandomForestClassifier method\), 1108](#)
[predict_proba\(\) \(evalml.pipelines.RandomForestRegressor method\), 1110](#)
[predict_proba\(\) \(evalml.pipelines.StackedEnsembleClassifier method\), 1124](#)
[predict_proba\(\) \(evalml.pipelines.StackedEnsembleRegressor method\), 1127](#)
[predict_proba\(\) \(evalml.pipelines.SVMClassifier method\), 1131](#)
[predict_proba\(\) \(evalml.pipelines.SVMRegressor method\), 1133](#)
[predict_proba\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesBinaryClassificationPipeline method\), 1018](#)
[predict_proba\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline method\), 1023](#)
[predict_proba\(\) \(evalml.pipelines.time_series_classification_pipeline.TimeSeriesMulticlassClassificationPipeline method\), 1148](#)
[predict_proba\(\) \(evalml.pipelines.XGBoostClassifier method\), 1157](#)
[predict_proba\(\) \(evalml.pipelines.XGBoostRegressor method\), 1160](#)
[print_deps\(\) \(in module evalml.utils.cli_utils\), 1188](#)
[print_info\(\) \(in module evalml.utils.cli_utils\), 1188](#)
[print_sys_info\(\) \(in module evalml.utils.cli_utils\), 1188](#)
[ProblemTypes \(class in evalml.problem_types\), 1176](#)
[ProblemTypes \(class in evalml.problem_types.problem_types\), 1172](#)
[propose\(\) \(evalml.tuners.grid_search_tuner.GridSearchTuner method\), 1178](#)
[propose\(\) \(evalml.tuners.GridSearchTuner method\), 1183](#)
[propose\(\) \(evalml.tuners.random_search_tuner.RandomSearchTuner method\), 1179](#)
[propose\(\) \(evalml.tuners.RandomSearchTuner method\), 1185](#)
[propose\(\) \(evalml.tuners.skopt_tuner.SKOptTuner method\), 1181](#)
[propose\(\) \(evalml.tuners.SKOptTuner method\), 1185](#)
[propose\(\) \(evalml.tuners.Tuner method\), 1186](#)
[propose\(\) \(evalml.tuners.tuner.Tuner method\), 1182](#)

R

[R2 \(class in evalml.objectives.standard_metrics\), 333](#)
[r2_callback\(\) \(in module evalml.callbacks\), 172](#)
[RandomForestClassifier \(class in evalml.pipelines\), 1106](#)
[RandomForestClassifier \(class in evalml.pipelines.components\), 934](#)
[RandomForestClassifier \(class in evalml.pipelines.components.estimators\), 626](#)
[RandomForestClassifier \(class in evalml.pipelines.components.estimators.classifiers\), 501](#)
[RandomForestClassifier \(class in evalml.pipelines.components.estimators.classifiers.rf_classifier\), 1140](#)
[RandomForestRegressor \(class in evalml.pipelines\), 1108](#)

[RandomForestRegressor](#) (class in [evalml.pipelines.components](#)), 937
[RandomForestRegressor](#) (class in [evalml.pipelines.components.estimators](#)), 628
[RandomForestRegressor](#) (class in [evalml.pipelines](#)), 1114
[RandomForestRegressor](#) (class in [evalml.pipelines.components.estimators.regressors](#)), 566
[RandomForestRegressor](#) (class in [evalml.pipelines.components.estimators.regressors.rf_regressor](#)), 534
[RandomSearchTuner](#) (class in [evalml.tuners](#)), 1184
[RandomSearchTuner](#) (class in [evalml.tuners.random_search_tuner](#)), 1178
[rankings\(\)](#) ([evalml.automl.automl_search.AutoMLSearch](#) property), 170
[rankings\(\)](#) ([evalml.AutoMLSearch](#) property), 180
[rankings\(\)](#) ([evalml.AutoMLSearch](#) property), 1202
[Recall](#) (class in [evalml.objectives](#)), 413
[Recall](#) (class in [evalml.objectives.standard_metrics](#)), 335
[RecallMacro](#) (class in [evalml.objectives](#)), 415
[RecallMacro](#) (class in [evalml.objectives.standard_metrics](#)), 337
[RecallMicro](#) (class in [evalml.objectives](#)), 416
[RecallMicro](#) (class in [evalml.objectives.standard_metrics](#)), 338
[RecallWeighted](#) (class in [evalml.objectives](#)), 418
[RecallWeighted](#) (class in [evalml.objectives.standard_metrics](#)), 340
[register\(\)](#) ([evalml.pipelines.components.component_base_meta.ComponentBaseMeta](#) method), 851
[register\(\)](#) ([evalml.pipelines.components.ComponentBaseMeta](#) method), 872
[register\(\)](#) ([evalml.pipelines.components.transformers.encoders.one_hot_encoder.OneHotEncoderMeta](#) method), 656
[register\(\)](#) ([evalml.pipelines.components.transformers.imputers.target_imputer.TargetImputerMeta](#) method), 695
[register\(\)](#) ([evalml.pipelines.pipeline_meta.PipelineBaseMeta](#) method), 1009
[register\(\)](#) ([evalml.pipelines.pipeline_meta.TimeSeriesPipelineBaseMeta](#) method), 1010
[register\(\)](#) ([evalml.utils.base_meta.BaseMeta](#) method), 1187
[RegressionObjective](#) (class in [evalml.objectives](#)), 420
[RegressionObjective](#) (class in [evalml.objectives.regression_objective](#)), 281
[RegressionPipeline](#) (class in [evalml.pipelines](#)), 1110
[RegressionPipeline](#) (class in [evalml.pipelines.regression_pipeline](#)), 1010
[results\(\)](#) ([evalml.automl.automl_search.AutoMLSearch](#) property), 170
[results\(\)](#) ([evalml.AutoMLSearch](#) property), 180
[results\(\)](#) ([evalml.AutoMLSearch](#) property), 1202
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines](#)), 1114
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components](#)), 939
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers](#)), 817
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection](#)), 677
[RFCClassifierSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection.rf_classifier](#)), 669
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines](#)), 1117
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components](#)), 942
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers](#)), 820
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection](#)), 680
[RFRegressorSelectFromModel](#) (class in [evalml.pipelines.components.transformers.feature_selection.rf_regressor](#)), 672
[roc_curve\(\)](#) (in module [evalml.model_understanding](#)), 264
[roc_curve\(\)](#) (in module [evalml.model_understanding](#)), 264
[RootMeanSquaredError](#) (class in [evalml.objectives](#)), 420
[RootMeanSquaredError](#) (class in [evalml.objectives.regression_metrics](#)), 342
[RootMeanSquaredLogError](#) (class in [evalml.objectives](#)), 423
[RootMeanSquaredLogError](#) (class in [evalml.objectives.standard_metrics](#)), 343

S

[safe_repr\(\)](#) (in module [evalml.utils](#)), 1197
[safe_repr\(\)](#) (in module [evalml.utils.gen_utils](#)), 1191
[SamplerBase](#) (class in [evalml.preprocessing.data_splitters](#)), 1165
[SamplerBase](#) (class in [evalml.preprocessing.data_splitters.sampler_base](#)), 1161
[save\(\)](#) ([evalml.automl.automl_search.AutoMLSearch](#) method), 170
[save\(\)](#) ([evalml.automl.AutoMLSearch](#) method), 180

`save()` (*evalml.AutoMLSearch* method), 1202
`save()` (*evalml.pipelines.ARIMARegressor* method), 1037
`save()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* method), 990
`save()` (*evalml.pipelines.BinaryClassificationPipeline* method), 1041
`save()` (*evalml.pipelines.CatBoostClassifier* method), 1044
`save()` (*evalml.pipelines.CatBoostRegressor* method), 1046
`save()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* method), 996
`save()` (*evalml.pipelines.ClassificationPipeline* method), 1050
`save()` (*evalml.pipelines.components.ARIMARegressor* method), 860
`save()` (*evalml.pipelines.components.BaselineClassifier* method), 862
`save()` (*evalml.pipelines.components.BaselineRegressor* method), 864
`save()` (*evalml.pipelines.components.CatBoostClassifier* method), 867
`save()` (*evalml.pipelines.components.CatBoostRegressor* method), 869
`save()` (*evalml.pipelines.components.component_base.ComponentBase* method), 850
`save()` (*evalml.pipelines.components.ComponentBase* method), 871
`save()` (*evalml.pipelines.components.DateTimeFeaturizer* method), 874
`save()` (*evalml.pipelines.components.DecisionTreeClassifier* method), 877
`save()` (*evalml.pipelines.components.DecisionTreeRegressor* method), 880
`save()` (*evalml.pipelines.components.DelayedFeatureTransformer* method), 882
`save()` (*evalml.pipelines.components.DFSTransformer* method), 884
`save()` (*evalml.pipelines.components.DropColumns* method), 886
`save()` (*evalml.pipelines.components.DropNullColumns* method), 888
`save()` (*evalml.pipelines.components.ElasticNetClassifier* method), 891
`save()` (*evalml.pipelines.components.ElasticNetRegressor* method), 894
`save()` (*evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsemble* method), 430
`save()` (*evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsembleClassifier* method), 433
`save()` (*evalml.pipelines.components.ensemble.stacked_ensemble.StackedEnsembleRegressor* method), 436
`save()` (*evalml.pipelines.components.ensemble.StackedEnsembleBase* method), 439
`save()` (*evalml.pipelines.components.ensemble.StackedEnsembleClassifier* method), 442
`save()` (*evalml.pipelines.components.ensemble.StackedEnsembleRegressor* method), 444
`save()` (*evalml.pipelines.components.Estimator* method), 896
`save()` (*evalml.pipelines.components.estimators.ARIMARegressor* method), 583
`save()` (*evalml.pipelines.components.estimators.BaselineClassifier* method), 585
`save()` (*evalml.pipelines.components.estimators.BaselineRegressor* method), 587
`save()` (*evalml.pipelines.components.estimators.CatBoostClassifier* method), 590
`save()` (*evalml.pipelines.components.estimators.CatBoostRegressor* method), 593
`save()` (*evalml.pipelines.components.estimators.classifiers.baseline_classifier.BaselineClassifier* method), 447
`save()` (*evalml.pipelines.components.estimators.classifiers.BaselineClassifier* method), 481
`save()` (*evalml.pipelines.components.estimators.classifiers.catboost_classifier.CatBoostClassifier* method), 450
`save()` (*evalml.pipelines.components.estimators.classifiers.CatBoostClassifier* method), 483
`save()` (*evalml.pipelines.components.estimators.classifiers.decision_tree_classifier.DecisionTreeClassifier* method), 453
`save()` (*evalml.pipelines.components.estimators.classifiers.DecisionTreeClassifier* method), 486
`save()` (*evalml.pipelines.components.estimators.classifiers.elasticnet_classifier.ElasticNetClassifier* method), 457
`save()` (*evalml.pipelines.components.estimators.classifiers.ElasticNetClassifier* method), 489
`save()` (*evalml.pipelines.components.estimators.classifiers.et_classifier.ExtraTreesClassifier* method), 460
`save()` (*evalml.pipelines.components.estimators.classifiers.ExtraTreesClassifier* method), 492
`save()` (*evalml.pipelines.components.estimators.classifiers.kneighbors_classifier.KNeighborsClassifier* method), 463
`save()` (*evalml.pipelines.components.estimators.classifiers.KNeighborsClassifier* method), 495
`save()` (*evalml.pipelines.components.estimators.classifiers.lightgbm_classifier.LightGBMClassifier* method), 466
`save()` (*evalml.pipelines.components.estimators.classifiers.LightGBMClassifier* method), 498
`save()` (*evalml.pipelines.components.estimators.classifiers.logistic_regression.LogisticRegression* method), 469
`save()` (*evalml.pipelines.components.estimators.classifiers.LogisticRegression* method), 501
`save()` (*evalml.pipelines.components.estimators.classifiers.RandomForestClassifier* method), 504
`save()` (*evalml.pipelines.components.estimators.classifiers.rf_classifier.RandomForestClassifier* method), 472
`save()` (*evalml.pipelines.components.estimators.classifiers.svm_classifier.SVC* method), 475

method), 475

save () (evalml.pipelines.components.estimators.classifiers.SVMClassifier), 506

save () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier), 478

save () (evalml.pipelines.components.estimators.classifiers.XGBoostClassifier), 509

save () (evalml.pipelines.components.estimators.DecisionTreeClassifier), 595

save () (evalml.pipelines.components.estimators.DecisionTreeRegressor), 598

save () (evalml.pipelines.components.estimators.ElasticNetClassifier), 601

save () (evalml.pipelines.components.estimators.ElasticNetRegressor), 604

save () (evalml.pipelines.components.estimators.Estimator), 606

save () (evalml.pipelines.components.estimators.Estimator), 579

save () (evalml.pipelines.components.estimators.ExtraTreesClassifier), 609

save () (evalml.pipelines.components.estimators.ExtraTreesRegressor), 612

save () (evalml.pipelines.components.estimators.KNeighborsClassifier), 615

save () (evalml.pipelines.components.estimators.LightGBMClassifier), 618

save () (evalml.pipelines.components.estimators.LightGBMRegressor), 620

save () (evalml.pipelines.components.estimators.LinearRegressor), 623

save () (evalml.pipelines.components.estimators.LogisticRegressionClassifier), 625

save () (evalml.pipelines.components.estimators.RandomForestClassifier), 628

save () (evalml.pipelines.components.estimators.RandomForestRegressor), 630

save () (evalml.pipelines.components.estimators.regressors.ariima), 512

save () (evalml.pipelines.components.estimators.regressors.Arimo), 548

save () (evalml.pipelines.components.estimators.regressors.baseline), 515

save () (evalml.pipelines.components.estimators.regressors.BaselineRegressor), 550

save () (evalml.pipelines.components.estimators.regressors.xgboost), 518

save () (evalml.pipelines.components.estimators.regressors.CatBoostRegressor), 553

save () (evalml.pipelines.components.estimators.regressors.decision_tree), 521

save () (evalml.pipelines.components.estimators.regressors.DecisionTreeRegressor), 555

save () (evalml.pipelines.components.estimators.regressors.elasticnet), 524

save () (evalml.pipelines.components.estimators.regressors.ElasticNetRegressor), 558

save () (evalml.pipelines.components.estimators.regressors.et_regressor), 527

save () (evalml.pipelines.components.estimators.regressors.ExtraTreesRegressor), 561

save () (evalml.pipelines.components.estimators.regressors.lightgbm_regressor), 530

save () (evalml.pipelines.components.estimators.regressors.LightGBMRegressor), 564

save () (evalml.pipelines.components.estimators.regressors.linear_regressor), 533

save () (evalml.pipelines.components.estimators.regressors.LinearRegressor), 566

save () (evalml.pipelines.components.estimators.regressors.RandomForestRegressor), 569

save () (evalml.pipelines.components.estimators.regressors.rf_regressor), 536

save () (evalml.pipelines.components.estimators.regressors.svm_regressor), 539

save () (evalml.pipelines.components.estimators.regressors.SVMRegressor), 571

save () (evalml.pipelines.components.estimators.regressors.time_series_baseline), 541

save () (evalml.pipelines.components.estimators.regressors.TimeSeriesBaselineRegressor), 574

save () (evalml.pipelines.components.estimators.regressors.xgboost_regressor), 544

save () (evalml.pipelines.components.estimators.regressors.XGBoostRegressor), 576

save () (evalml.pipelines.components.estimators.SVMClassifier), 633

save () (evalml.pipelines.components.estimators.SVMRegressor), 635

save () (evalml.pipelines.components.estimators.TimeSeriesBaselineEstimator), 637

save () (evalml.pipelines.components.estimators.XGBoostClassifier), 640

save () (evalml.pipelines.components.estimators.XGBoostRegressor), 642

save () (evalml.pipelines.components.estimators.ExtraTreesClassifier), 899

save () (evalml.pipelines.components.ExtraTreesRegressor), 902

save () (evalml.pipelines.components.FeatureSelector), 904

save () (evalml.pipelines.components.Imputer), 906

save () (evalml.pipelines.components.LightGBMClassifier), 912

save () (evalml.pipelines.components.LightGBMRegressor), 912

method), 915

save () (evalml.pipelines.components.LinearDiscriminantAnalysis), 917

save () (evalml.pipelines.components.LinearRegressor), 920

save () (evalml.pipelines.components.LogisticRegressionClassifier), 923

save () (evalml.pipelines.components.LSA), 924

save () (evalml.pipelines.components.OneHotEncoder), 927

save () (evalml.pipelines.components.PCA), 929

save () (evalml.pipelines.components.PerColumnImputer), 932

save () (evalml.pipelines.components.PolynomialDetrender), 934

save () (evalml.pipelines.components.RandomForestClassifier), 937

save () (evalml.pipelines.components.RandomForestRegressor), 939

save () (evalml.pipelines.components.RFClassifierSelectFromModel), 941

save () (evalml.pipelines.components.RFRegressorSelectFromModel), 944

save () (evalml.pipelines.components.SelectColumns), 946

save () (evalml.pipelines.components.SimpleImputer), 948

save () (evalml.pipelines.components.SMOTENCSampler), 951

save () (evalml.pipelines.components.SMOTENSampler), 953

save () (evalml.pipelines.components.SMOTESampler), 955

save () (evalml.pipelines.components.StackedEnsembleClassifier), 958

save () (evalml.pipelines.components.StackedEnsembleRegressor), 960

save () (evalml.pipelines.components.StandardScaler), 962

save () (evalml.pipelines.components.SVMClassifier), 965

save () (evalml.pipelines.components.SVMRegressor), 967

save () (evalml.pipelines.components.TargetEncoder), 969

save () (evalml.pipelines.components.TargetImputer), 971

save () (evalml.pipelines.components.TextFeaturizer), 973

save () (evalml.pipelines.components.TimeSeriesBaselineEstimator), 976

save () (evalml.pipelines.components.Transformer), 978

save () (evalml.pipelines.components.transformers.column_selectors.ColumnSelector), 776

save () (evalml.pipelines.components.transformers.column_selectors.DropColumns), 778

save () (evalml.pipelines.components.transformers.column_selectors.SelectColumns), 780

save () (evalml.pipelines.components.transformers.DateTimeFeaturizer), 789

save () (evalml.pipelines.components.transformers.DelayedFeatureTransformer), 791

save () (evalml.pipelines.components.transformers.DFSTransformer), 793

save () (evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimImputer), 645

save () (evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimRegressor), 650

save () (evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimClassifier), 652

save () (evalml.pipelines.components.transformers.dimensionality_reduction.ReducedDimRegressor), 647

save () (evalml.pipelines.components.transformers.DropColumns), 795

save () (evalml.pipelines.components.transformers.DropNullColumns), 798

save () (evalml.pipelines.components.transformers.encoders.onehot_encoder.OneHotEncoder), 655

save () (evalml.pipelines.components.transformers.encoders.OneHotEncoder), 662

save () (evalml.pipelines.components.transformers.encoders.target_encoder.TargetEncoder), 659

save () (evalml.pipelines.components.transformers.encoders.TargetEncoder), 665

save () (evalml.pipelines.components.transformers.feature_selection.FeatureSelector), 668

save () (evalml.pipelines.components.transformers.feature_selection.FeatureSelector), 677

save () (evalml.pipelines.components.transformers.feature_selection.rf_classifier.RFClassifier), 671

save () (evalml.pipelines.components.transformers.feature_selection.rf_regressor.RFRegressor), 674

save () (evalml.pipelines.components.transformers.feature_selection.RFC), 680

save () (evalml.pipelines.components.transformers.feature_selection.RFR), 683

save () (evalml.pipelines.components.transformers.FeatureSelector), 800

save () (evalml.pipelines.components.transformers.Imputer), 802

save () (evalml.pipelines.components.transformers.imputers.Imputer), 697

save () (evalml.pipelines.components.transformers.imputers.imputer.Imputer), 686

save () (evalml.pipelines.components.transformers.imputers.per_column_imputer.PerColumnImputer), 697

method), 688

save () (evalml.pipelines.components.transformers.imputers.PerColumnImputer), 699

save () (evalml.pipelines.components.transformers.imputers.SimpleImputer), 691

save () (evalml.pipelines.components.transformers.imputers.TargetEncoder), 702

save () (evalml.pipelines.components.transformers.imputers.TargetEncoder), 694

save () (evalml.pipelines.components.transformers.imputers.TargetEncoder), 704

save () (evalml.pipelines.components.transformers.LinearDiscriminantAnalysis), 804

save () (evalml.pipelines.components.transformers.LSA), 806

save () (evalml.pipelines.components.transformers.OneHotEncoder), 809

save () (evalml.pipelines.components.transformers.PCA), 812

save () (evalml.pipelines.components.transformers.PerColumnImputer), 814

save () (evalml.pipelines.components.transformers.PolynomialDegreeReducer), 816

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 707

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 728

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 710

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 731

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 733

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 712

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 735

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 715

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 737

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 717

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 720

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 740

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 722

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 725

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 742

save () (evalml.pipelines.components.transformers.preprocessing.DelayedFeatureTransformer), 744

method), 744

save () (evalml.pipelines.components.transformers.RFClassifierSelectFromModel), 819

save () (evalml.pipelines.components.transformers.RFRegressorSelectFromModel), 822

save () (evalml.pipelines.components.transformers.samplers.base_sampler), 747

save () (evalml.pipelines.components.transformers.samplers.base_sampler), 749

save () (evalml.pipelines.components.transformers.samplers.oversampler), 752

save () (evalml.pipelines.components.transformers.samplers.oversampler), 754

save () (evalml.pipelines.components.transformers.samplers.oversampler), 756

save () (evalml.pipelines.components.transformers.samplers.SMOTE), 762

save () (evalml.pipelines.components.transformers.samplers.SMOTENS), 764

save () (evalml.pipelines.components.transformers.samplers.SMOTESampler), 766

save () (evalml.pipelines.components.transformers.samplers.Undersampler), 769

save () (evalml.pipelines.components.transformers.samplers.undersampler), 759

save () (evalml.pipelines.components.transformers.scalers.standard_scaler), 771

save () (evalml.pipelines.components.transformers.DelayedFeatureTransformer), 774

save () (evalml.pipelines.components.transformers.SelectColumns), 824

save () (evalml.pipelines.components.transformers.SimpleImputer), 826

save () (evalml.pipelines.components.transformers.SMOTE), 829

save () (evalml.pipelines.components.transformers.SMOTENS), 831

save () (evalml.pipelines.components.transformers.SMOTESampler), 833

save () (evalml.pipelines.components.transformers.StandardScaler), 835

save () (evalml.pipelines.components.transformers.TargetEncoder), 838

save () (evalml.pipelines.components.transformers.TargetImputer), 840

save () (evalml.pipelines.components.transformers.TextFeaturizer), 842

save () (evalml.pipelines.components.transformers.Transformer), 844

save () (evalml.pipelines.components.transformers.transformer.TargetTransformer), 783

save () (evalml.pipelines.components.transformers.transformer.Transformer), 785

save () (evalml.pipelines.components.transformers.Undersampler), 785

- method*), 847
- `save()` (*evalml.pipelines.components.Undersampler method*), 981
- `save()` (*evalml.pipelines.components.XGBoostClassifier method*), 983
- `save()` (*evalml.pipelines.components.XGBoostRegressor method*), 986
- `save()` (*evalml.pipelines.DecisionTreeClassifier method*), 1057
- `save()` (*evalml.pipelines.DecisionTreeRegressor method*), 1059
- `save()` (*evalml.pipelines.DelayedFeatureTransformer method*), 1061
- `save()` (*evalml.pipelines.DFSTransformer method*), 1064
- `save()` (*evalml.pipelines.ElasticNetClassifier method*), 1067
- `save()` (*evalml.pipelines.ElasticNetRegressor method*), 1069
- `save()` (*evalml.pipelines.Estimator method*), 1071
- `save()` (*evalml.pipelines.ExtraTreesClassifier method*), 1074
- `save()` (*evalml.pipelines.ExtraTreesRegressor method*), 1077
- `save()` (*evalml.pipelines.FeatureSelector method*), 1079
- `save()` (*evalml.pipelines.KNeighborsClassifier method*), 1082
- `save()` (*evalml.pipelines.LightGBMClassifier method*), 1085
- `save()` (*evalml.pipelines.LightGBMRegressor method*), 1088
- `save()` (*evalml.pipelines.LinearRegressor method*), 1090
- `save()` (*evalml.pipelines.LogisticRegressionClassifier method*), 1093
- `save()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline method*), 1004
- `save()` (*evalml.pipelines.MulticlassClassificationPipeline method*), 1096
- `save()` (*evalml.pipelines.OneHotEncoder method*), 1099
- `save()` (*evalml.pipelines.PerColumnImputer method*), 1102
- `save()` (*evalml.pipelines.pipeline_base.PipelineBase method*), 1008
- `save()` (*evalml.pipelines.PipelineBase method*), 1105
- `save()` (*evalml.pipelines.RandomForestClassifier method*), 1108
- `save()` (*evalml.pipelines.RandomForestRegressor method*), 1110
- `save()` (*evalml.pipelines.regression_pipeline.RegressionPipeline method*), 1014
- `save()` (*evalml.pipelines.RegressionPipeline method*), 1114
- `save()` (*evalml.pipelines.RFClassifierSelectFromModel method*), 1116
- `save()` (*evalml.pipelines.RFRegressorSelectFromModel method*), 1119
- `save()` (*evalml.pipelines.SimpleImputer method*), 1121
- `save()` (*evalml.pipelines.StackedEnsembleClassifier method*), 1124
- `save()` (*evalml.pipelines.StackedEnsembleRegressor method*), 1127
- `save()` (*evalml.pipelines.StandardScaler method*), 1129
- `save()` (*evalml.pipelines.SVMClassifier method*), 1131
- `save()` (*evalml.pipelines.SVMRegressor method*), 1134
- `save()` (*evalml.pipelines.TargetEncoder method*), 1136
- `save()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), 1019
- `save()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), 1023
- `save()` (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline method*), 1027
- `save()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline method*), 1032
- `save()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline method*), 1140
- `save()` (*evalml.pipelines.TimeSeriesClassificationPipeline method*), 1144
- `save()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline method*), 1148
- `save()` (*evalml.pipelines.TimeSeriesRegressionPipeline method*), 1152
- `save()` (*evalml.pipelines.Transformer method*), 1155
- `save()` (*evalml.pipelines.XGBoostClassifier method*), 1157
- `save()` (*evalml.pipelines.XGBoostRegressor method*), 1160
- `save_plot()` (*in module evalml.utils.gen_utils*), 1192
- `scikit_learn_wrapped_estimator()` (*in module evalml.pipelines.components.utils*), 852
- `score()` (*evalml.objectives.AccuracyBinary method*), 352
- `score()` (*evalml.objectives.AccuracyMulticlass method*), 353
- `score()` (*evalml.objectives.AUC method*), 355
- `score()` (*evalml.objectives.AUCMacro method*), 357
- `score()` (*evalml.objectives.AUCMicro method*), 359
- `score()` (*evalml.objectives.AUCWeighted method*), 360
- `score()` (*evalml.objectives.BalancedAccuracyBinary method*), 362
- `score()` (*evalml.objectives.BalancedAccuracyMulticlass method*), 364
- `score()` (*evalml.objectives.binary_classification_objective.BinaryClassificationObjective method*), 267

`score()` (`evalml.objectives.BinaryClassificationObjective` method), 366
`score()` (`evalml.objectives.cost_benefit_matrix.CostBenefitMatrix` method), 270
`score()` (`evalml.objectives.CostBenefitMatrix` method), 369
`score()` (`evalml.objectives.ExpVariance` method), 371
`score()` (`evalml.objectives.F1` method), 373
`score()` (`evalml.objectives.F1Macro` method), 374
`score()` (`evalml.objectives.F1Micro` method), 376
`score()` (`evalml.objectives.F1Weighted` method), 378
`score()` (`evalml.objectives.fraud_cost.FraudCost` method), 273
`score()` (`evalml.objectives.FraudCost` method), 380
`score()` (`evalml.objectives.lead_scoring.LeadScoring` method), 275
`score()` (`evalml.objectives.LeadScoring` method), 383
`score()` (`evalml.objectives.LogLossBinary` method), 385
`score()` (`evalml.objectives.LogLossMulticlass` method), 387
`score()` (`evalml.objectives.MAE` method), 388
`score()` (`evalml.objectives.MAPE` method), 390
`score()` (`evalml.objectives.MaxError` method), 391
`score()` (`evalml.objectives.MCCBinary` method), 394
`score()` (`evalml.objectives.MCCMulticlass` method), 395
`score()` (`evalml.objectives.MeanSquaredLogError` method), 397
`score()` (`evalml.objectives.MedianAE` method), 398
`score()` (`evalml.objectives.MSE` method), 400
`score()` (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` method), 277
`score()` (`evalml.objectives.MulticlassClassificationObjective` method), 402
`score()` (`evalml.objectives.objective_base.ObjectiveBase` method), 280
`score()` (`evalml.objectives.ObjectiveBase` method), 404
`score()` (`evalml.objectives.Precision` method), 406
`score()` (`evalml.objectives.PrecisionMacro` method), 408
`score()` (`evalml.objectives.PrecisionMicro` method), 409
`score()` (`evalml.objectives.PrecisionWeighted` method), 411
`score()` (`evalml.objectives.R2` method), 412
`score()` (`evalml.objectives.Recall` method), 414
`score()` (`evalml.objectives.RecallMacro` method), 416
`score()` (`evalml.objectives.RecallMicro` method), 418
`score()` (`evalml.objectives.RecallWeighted` method), 419
`score()` (`evalml.objectives.regression_objective.RegressionObjective` method), 282
`score()` (`evalml.objectives.RegressionObjective` method), 421
`score()` (`evalml.objectives.RootMeanSquaredError` method), 423
`score()` (`evalml.objectives.RootMeanSquaredLogError` method), 424
`score()` (`evalml.objectives.sensitivity_low_alert.SensitivityLowAlert` method), 285
`score()` (`evalml.objectives.SensitivityLowAlert` method), 426
`score()` (`evalml.objectives.standard_metrics.AccuracyBinary` method), 288
`score()` (`evalml.objectives.standard_metrics.AccuracyMulticlass` method), 290
`score()` (`evalml.objectives.standard_metrics.AUC` method), 292
`score()` (`evalml.objectives.standard_metrics.AUCMacro` method), 293
`score()` (`evalml.objectives.standard_metrics.AUCMicro` method), 295
`score()` (`evalml.objectives.standard_metrics.AUCWeighted` method), 297
`score()` (`evalml.objectives.standard_metrics.BalancedAccuracyBinary` method), 299
`score()` (`evalml.objectives.standard_metrics.BalancedAccuracyMulticlass` method), 300
`score()` (`evalml.objectives.standard_metrics.ExpVariance` method), 302
`score()` (`evalml.objectives.standard_metrics.F1` method), 304
`score()` (`evalml.objectives.standard_metrics.F1Macro` method), 306
`score()` (`evalml.objectives.standard_metrics.F1Micro` method), 307
`score()` (`evalml.objectives.standard_metrics.F1Weighted` method), 309
`score()` (`evalml.objectives.standard_metrics.LogLossBinary` method), 311
`score()` (`evalml.objectives.standard_metrics.LogLossMulticlass` method), 313
`score()` (`evalml.objectives.standard_metrics.MAE` method), 314
`score()` (`evalml.objectives.standard_metrics.MAPE` method), 316
`score()` (`evalml.objectives.standard_metrics.MaxError` method), 317
`score()` (`evalml.objectives.standard_metrics.MCCBinary` method), 319
`score()` (`evalml.objectives.standard_metrics.MCCMulticlass` method), 321
`score()` (`evalml.objectives.standard_metrics.MeanSquaredLogError` method), 323
`score()` (`evalml.objectives.standard_metrics.MedianAE` method), 324

`score()` (`evalml.objectives.standard_metrics.MSE` method), 1027
`score()` (`evalml.objectives.standard_metrics.Precision` method), 326
`score()` (`evalml.objectives.standard_metrics.PrecisionMacro` method), 328
`score()` (`evalml.objectives.standard_metrics.PrecisionMacro` method), 330
`score()` (`evalml.objectives.standard_metrics.PrecisionMicro` method), 331
`score()` (`evalml.objectives.standard_metrics.PrecisionWeighted` method), 333
`score()` (`evalml.objectives.standard_metrics.R2` method), 334
`score()` (`evalml.objectives.standard_metrics.Recall` method), 336
`score()` (`evalml.objectives.standard_metrics.RecallMacro` method), 338
`score()` (`evalml.objectives.standard_metrics.RecallMicro` method), 340
`score()` (`evalml.objectives.standard_metrics.RecallWeighted` method), 341
`score()` (`evalml.objectives.standard_metrics.RootMeanSquaredError` method), 343
`score()` (`evalml.objectives.standard_metrics.RootMeanSquaredLogError` method), 344
`score()` (`evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective` method), 346
`score()` (`evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline` method), 990
`score()` (`evalml.pipelines.BinaryClassificationPipeline` method), 1041
`score()` (`evalml.pipelines.classification_pipeline.ClassificationPipeline` method), 996
`score()` (`evalml.pipelines.ClassificationPipeline` method), 1050
`score()` (`evalml.pipelines.components.utils.WrappedSKClassifier` method), 853
`score()` (`evalml.pipelines.components.utils.WrappedSKRegressor` method), 854
`score()` (`evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline` method), 1004
`score()` (`evalml.pipelines.MulticlassClassificationPipeline` method), 1097
`score()` (`evalml.pipelines.pipeline_base.PipelineBase` method), 1008
`score()` (`evalml.pipelines.PipelineBase` method), 1106
`score()` (`evalml.pipelines.regression_pipeline.RegressionPipeline` method), 1014
`score()` (`evalml.pipelines.RegressionPipeline` method), 1114
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1019
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1023
`score()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1027
`score()` (`evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline` method), 1032
`score()` (`evalml.pipelines.TimeSeriesBinaryClassificationPipeline` method), 1140
`score()` (`evalml.pipelines.TimeSeriesClassificationPipeline` method), 1144
`score()` (`evalml.pipelines.TimeSeriesMulticlassClassificationPipeline` method), 1149
`score()` (`evalml.pipelines.TimeSeriesRegressionPipeline` method), 1153
`score_needs_proba()` (`evalml.objectives.binary_classification_objective.BinaryClassificationObjective` property), 267
`score_needs_proba()` (`evalml.objectives.BinaryClassificationObjective` property), 367
`score_needs_proba()` (`evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective` property), 278
`score_needs_proba()` (`evalml.objectives.MulticlassClassificationObjective` property), 402
`score_needs_proba()` (`evalml.objectives.regression_objective.RegressionObjective` property), 280
`score_needs_proba()` (`evalml.objectives.ObjectiveBase` property), 404
`score_needs_proba()` (`evalml.objectives.RegressionObjective` property), 421
`score_needs_proba()` (`evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective` property), 347
`score_pipelines()` (`evalml.pipelines.MulticlassClassificationPipeline` method), 157
`score_pipelines()` (`evalml.pipelines.MulticlassClassificationPipeline` method), 170
`score_pipelines()` (`evalml.pipelines.PipelineBase` method), 180
`score_pipelines()` (`evalml.pipelines.RegressionPipeline` method), 1202
`search()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 180
`search()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 1202
`search()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 182
`search()` (`evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline` method), 182

762	evalml.pipelines.components.ensemble.stacked_ensemble_classifier
SMOTENSampler (class in	431
evalml.pipelines.components.transformers.sampled_syntetic_sampler),	1124
752	evalml.pipelines), 1124
SMOTESampler (class in	StackedEnsembleRegressor (class in
evalml.pipelines.components), 953	evalml.pipelines.components), 958
SMOTESampler (class in	StackedEnsembleRegressor (class in
evalml.pipelines.components.transformers),	evalml.pipelines.components.ensemble),
831	442
SMOTESampler (class in	StackedEnsembleRegressor (class in
evalml.pipelines.components.transformers.samplers),	evalml.pipelines.components.ensemble.stacked_ensemble_regressor
765	434
SMOTESampler (class in	StandardScaler (class in evalml.pipelines), 1127
evalml.pipelines.components.transformers.sampled_syntetic_sampler),	StandardScaler (class in
754	evalml.pipelines.components), 961
sparsity_score() (evalml.data_checks.sparsity_data_check.SparsityDataCheck	(class in
static method), 203	evalml.pipelines.components.transformers),
sparsity_score() (evalml.data_checks.SparsityDataCheck	833
static method), 222	StandardScaler (class in
SparsityDataCheck (class in evalml.data_checks),	evalml.pipelines.components.transformers.scalers),
222	772
SparsityDataCheck (class in	StandardScaler (class in
evalml.data_checks.sparsity_data_check),	evalml.pipelines.components.transformers.scalers.standard_scaler
203	770
split() (evalml.preprocessing.data_splitters.time_series_split.TimeSeriesSplit	evalml.automl.engine.cf_engine.CFClient
method), 1162	method), 151
split() (evalml.preprocessing.data_splitters.TimeSeriesSplit	submit_evaluation_job()
method), 1166	(evalml.automl.engine.cf_engine.CFEngine
split() (evalml.preprocessing.data_splitters.training_validation_split.TrainingValidationSplit	method), 159
method), 1164	submit_evaluation_job()
split() (evalml.preprocessing.data_splitters.TrainingValidationSplit	(evalml.automl.engine.CFEngine method),
method), 1166	160
split() (evalml.preprocessing.TimeSeriesSplit	submit_evaluation_job()
method), 1170	(evalml.automl.engine.dask_engine.DaskEngine
split() (evalml.preprocessing.TrainingValidationSplit	method), 154
method), 1171	submit_evaluation_job()
split_data() (in module evalml.preprocessing),	(evalml.automl.engine.DaskEngine method),
1169	162
split_data() (in module evalml.preprocessing.utils),	submit_evaluation_job()
1168	(evalml.automl.engine.engine_base.EngineBase
StackedEnsembleBase (class in	method), 156
evalml.pipelines.components.ensemble),	submit_evaluation_job()
437	(evalml.automl.engine.EngineBase method),
StackedEnsembleBase (class in	163
evalml.pipelines.components.ensemble.stacked_ensemble_base),	submit_evaluation_job()
428	(evalml.automl.engine.sequential_engine.SequentialEngine
StackedEnsembleClassifier (class in	method), 159
evalml.pipelines), 1122	submit_evaluation_job()
StackedEnsembleClassifier (class in	(evalml.automl.engine.SequentialEngine
evalml.pipelines.components), 955	method), 164
StackedEnsembleClassifier (class in	submit_evaluation_job()
evalml.pipelines.components.ensemble),	(evalml.automl.EngineBase method), 181
439	submit_evaluation_job()
StackedEnsembleClassifier (class in	(evalml.automl.SequentialEngine method),

183

`submit_scoring_job()`
(*evalml.automl.engine.cf_engine.CFEngine* method), 152

`submit_scoring_job()`
(*evalml.automl.engine.CFEngine* method), 161

`submit_scoring_job()`
(*evalml.automl.engine.dask_engine.DaskEngine* method), 154

`submit_scoring_job()`
(*evalml.automl.engine.DaskEngine* method), 162

`submit_scoring_job()`
(*evalml.automl.engine.engine_base.EngineBase* method), 156

`submit_scoring_job()`
(*evalml.automl.engine.EngineBase* method), 163

`submit_scoring_job()`
(*evalml.automl.engine.sequential_engine.SequentialEngine* method), 159

`submit_scoring_job()`
(*evalml.automl.engine.SequentialEngine* method), 164

`submit_scoring_job()`
(*evalml.automl.EngineBase* method), 181

`submit_scoring_job()`
(*evalml.automl.SequentialEngine* method), 183

`submit_training_job()`
(*evalml.automl.engine.cf_engine.CFEngine* method), 152

`submit_training_job()`
(*evalml.automl.engine.CFEngine* method), 161

`submit_training_job()`
(*evalml.automl.engine.dask_engine.DaskEngine* method), 155

`submit_training_job()`
(*evalml.automl.engine.DaskEngine* method), 162

`submit_training_job()`
(*evalml.automl.engine.engine_base.EngineBase* method), 156

`submit_training_job()`
(*evalml.automl.engine.EngineBase* method), 163

`submit_training_job()`
(*evalml.automl.engine.sequential_engine.SequentialEngine* method), 159

`submit_training_job()`
(*evalml.automl.engine.SequentialEngine* method), 164

`submit_training_job()`
(*evalml.automl.EngineBase* method), 181

`submit_training_job()`
(*evalml.automl.SequentialEngine* method), 183

`summary()` (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* property), 991

`summary()` (*evalml.pipelines.BinaryClassificationPipeline* property), 1042

`summary()` (*evalml.pipelines.classification_pipeline.ClassificationPipeline* property), 996

`summary()` (*evalml.pipelines.ClassificationPipeline* property), 1051

`summary()` (*evalml.pipelines.multiclass_classification_pipeline.MulticlassClassificationPipeline* property), 1004

`summary()` (*evalml.pipelines.MulticlassClassificationPipeline* property), 1097

`summary()` (*evalml.pipelines.pipeline_base.PipelineBase* property), 1008

`summary()` (*evalml.pipelines.PipelineBase* property), 1106

`summary()` (*evalml.pipelines.regression_pipeline.RegressionPipeline* property), 1014

`summary()` (*evalml.pipelines.RegressionPipeline* property), 1114

`summary()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* property), 1019

`summary()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* property), 1023

`summary()` (*evalml.pipelines.time_series_classification_pipeline.TimeSeriesClassificationPipeline* property), 1028

`summary()` (*evalml.pipelines.time_series_regression_pipeline.TimeSeriesRegressionPipeline* property), 1032

`summary()` (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* property), 1140

`summary()` (*evalml.pipelines.TimeSeriesClassificationPipeline* property), 1145

`summary()` (*evalml.pipelines.TimeSeriesMulticlassClassificationPipeline* property), 1149

`summary()` (*evalml.pipelines.TimeSeriesRegressionPipeline* property), 1153

`supported_problem_types()`
(*evalml.pipelines.components.ensemble.stacked_ensemble_base.StackedEnsembleBase* property), 430

`supported_problem_types()`
(*evalml.pipelines.components.ensemble.StackedEnsembleBase* property), 439

`supported_problem_types()`
(*evalml.pipelines.components.Estimator* property), 896

`supported_problem_types()`
(*evalml.pipelines.components.estimators.Estimator* property), 606

`supported_problem_types()`

(*evalml.pipelines.components.estimators.estimator* *Estimator* *property*), 579

supported_problem_types() (*evalml.pipelines.Estimator* *property*), 1071

SVMClassifier (*class* *in evalml.pipelines*), 1129

SVMClassifier (*class* *in evalml.pipelines.components*), 963

SVMClassifier (*class* *in evalml.pipelines.components.estimators*), 630

SVMClassifier (*class* *in evalml.pipelines.components.estimators.classifiers*), 504

SVMClassifier (*class* *in evalml.pipelines.components.estimators.classifiers.target_classifier*), 473

SVMRegressor (*class* *in evalml.pipelines*), 1131

SVMRegressor (*class* *in evalml.pipelines.components*), 965

SVMRegressor (*class* *in evalml.pipelines.components.estimators*), 633

SVMRegressor (*class* *in evalml.pipelines.components.estimators.regressors*), 569

SVMRegressor (*class* *in evalml.pipelines.components.estimators.regressors.svm_regressor*), 537

T

t_sne() (*in module evalml.model_understanding*), 264

t_sne() (*in module evalml.model_understanding.graphs*), 251

target_distribution() (*in module evalml.preprocessing*), 1170

target_distribution() (*in module evalml.preprocessing.utils*), 1168

TargetEncoder (*class* *in evalml.pipelines*), 1134

TargetEncoder (*class* *in evalml.pipelines.components*), 967

TargetEncoder (*class* *in evalml.pipelines.components.transformers*), 835

TargetEncoder (*class* *in evalml.pipelines.components.transformers.encoders*), 663

TargetEncoder (*class* *in evalml.pipelines.components.transformers.encoders.target_encoder*), 657

TargetImputer (*class* *in evalml.pipelines.components*), 970

TargetImputer (*class* *in evalml.pipelines.components.transformers*), 838

TargetImputer (*class* *in evalml.pipelines.components.transformers.imputers*), 702

TargetImputer (*class* *in evalml.pipelines.components.transformers.imputers.target_imputer*), 692

TargetImputerMeta (*class* *in evalml.pipelines.components.transformers.imputers.target_imputer*), 694

TargetLeakageDataCheck (*class* *in evalml.data_checks*), 223

TargetLeakageDataCheck (*class* *in evalml.data_checks.target_leakage_data_check*), 205

TargetTransformer (*class* *in evalml.pipelines.components.transformers.transformer*), 781

TextFeaturizer (*class* *in evalml.pipelines.components*), 972

TextFeaturizer (*class* *in evalml.pipelines.components.transformers*), 840

TextFeaturizer (*class* *in evalml.pipelines.components.transformers.preprocessing*), 740

TextFeaturizer (*class* *in evalml.pipelines.components.transformers.preprocessing.text_featurizer*), 721

TextTransformer (*class* *in evalml.pipelines.components.transformers.preprocessing*), 742

TextTransformer (*class* *in evalml.pipelines.components.transformers.preprocessing.text_transformer*), 723

threshold() (*evalml.pipelines.binary_classification_pipeline.BinaryClassificationPipeline* *property*), 991

threshold() (*evalml.pipelines.binary_classification_pipeline_mixin.BinaryClassificationPipelineMixin* *property*), 992

threshold() (*evalml.pipelines.BinaryClassificationPipeline* *property*), 1042

threshold() (*evalml.pipelines.time_series_classification_pipelines.TimeSeriesClassificationPipeline* *property*), 1019

threshold() (*evalml.pipelines.TimeSeriesBinaryClassificationPipeline* *property*), 1141

Time_elapsed() (*in module evalml.utils.logger*), 1193

TimeSeriesBaselineEstimator (*class* *in evalml.pipelines.components*), 974

TimeSeriesBaselineEstimator (*class* *in evalml.pipelines.components.estimators*), 635

TimeSeriesBaselineEstimator (*class* *in evalml.pipelines.components.estimators.regressors*), 571

TimeSeriesBaselineEstimator (*class* *in*

`evalml.pipelines.components.estimators.regressor_time_series_baseline_estimator`, (in module `evalml.automl.engine`), 164

`TimeSeriesBinaryClassificationPipeline` `train_pipeline()` (in module `evalml.automl.engine.engine_base`), 158

`TimeSeriesBinaryClassificationPipeline` `train_pipeline()` (class in `evalml.pipelines`), 1136

`TimeSeriesBinaryClassificationPipeline` `train_pipeline()` (class in `evalml.pipelines.time_series_classification_pipelines`), 1015

`TimeSeriesClassificationPipeline` `train_pipeline()` (class in `evalml.pipelines`), 1141

`TimeSeriesClassificationPipeline` `train_pipeline()` (class in `evalml.pipelines.time_series_classification_pipelines`), 1019

`TimeSeriesMulticlassClassificationPipeline` `TrainingValidationSplit` (class in `evalml.preprocessing`), 1171

`TimeSeriesMulticlassClassificationPipeline` `TrainingValidationSplit` (class in `evalml.preprocessing.data_splitters`), 1166

`TimeSeriesMulticlassClassificationPipeline` `TrainingValidationSplit` (class in `evalml.preprocessing.data_splitters.training_validation_split`), 1023

`TimeSeriesPipelineBaseMeta` (class in `evalml.pipelines.pipeline_meta`), 1009

`TimeSeriesRegressionObjective` (class in `evalml.objectives.time_series_regression_objective`), 345

`TimeSeriesRegressionPipeline` (class in `evalml.pipelines`), 1149

`TimeSeriesRegressionPipeline` (class in `evalml.pipelines.time_series_regression_pipeline`), 1028

`TimeSeriesSplit` (class in `evalml.preprocessing`), 1170

`TimeSeriesSplit` (class in `evalml.preprocessing.data_splitters`), 1165

`TimeSeriesSplit` (class in `evalml.preprocessing.data_splitters.time_series_split`), 1162

`to_dict()` (`evalml.data_checks.data_check_action.DataCheckAction`), 186

`to_dict()` (`evalml.data_checks.data_check_message.DataCheckError`), 188

`to_dict()` (`evalml.data_checks.data_check_message.DataCheckMessage`), 188

`to_dict()` (`evalml.data_checks.data_check_message.DataCheckWarning`), 189

`to_dict()` (`evalml.data_checks.DataCheckAction`), 211

`to_dict()` (`evalml.data_checks.DataCheckError`), 212

`to_dict()` (`evalml.data_checks.DataCheckMessage`), 212

`to_dict()` (`evalml.data_checks.DataCheckWarning`), 215

`train_and_score_pipeline()` (in module `evalml.automl.engine`), 164

`train_and_score_pipeline()` (in module `evalml.automl.engine.engine_base`), 158

`transform()` (`evalml.pipelines.components.DateFeatureTransformer`), 874

`transform()` (`evalml.pipelines.components.DelayedFeatureTransformer`), 882

`transform()` (`evalml.pipelines.components.DFSTransformer`), 884

`transform()` (`evalml.pipelines.components.DropColumns`), 886

`transform()` (`evalml.pipelines.components.DropNullColumns`), 888

`transform()` (`evalml.pipelines.components.FeatureSelector`), 904

`transform()` (`evalml.pipelines.components.Imputer`), 906

`transform()` (`evalml.pipelines.components.LinearDiscriminantAnalysis`), 917

`transform()` (`evalml.pipelines.components.LSA`), 924

`transform()` (`evalml.pipelines.components.OneHotEncoder`), 927

`transform()` (`evalml.pipelines.components.PCA`), 930

`transform()` (`evalml.pipelines.components.PerColumnImputer`), 932

`transform()` (`evalml.pipelines.components.PolynomialDetrender`), 934

`transform()` (`evalml.pipelines.components.RFClassifierSelectFromModel`), 942

`transform()` (`evalml.pipelines.components.RFRegressorSelectFromModel`), 944

`transform()` (`evalml.pipelines.components.SelectColumns`), 946

`transform()` (`evalml.pipelines.components.SimpleImputer`), 949

`transform()` (`evalml.pipelines.components.SMOTENC Sampler`), 951

`transform()` (`evalml.pipelines.components.SMOTENSMOTE`), 953

`transform()` (`evalml.pipelines.components.SMOTESampler`), 955

`transform()` (`evalml.pipelines.components.StandardScaler`), 962

`transform()` (`evalml.pipelines.components.TargetEncoder`), 969

`transform()` (`evalml.pipelines.components.TargetImputer`), 972

`transform()` (`evalml.pipelines.components.TextFeaturizer`), 974

`transform()` (`evalml.pipelines.components.Transformer`), 978

`transform()` (`evalml.pipelines.components.transformers.ColumnSelector`), 776

`transform()` (`evalml.pipelines.components.transformers.ColumnSelector`), 778

`transform()` (`evalml.pipelines.components.transformers.ColumnSelector`), 780

`transform()` (`evalml.pipelines.components.transformers.DateTimeFeaturizer`), 789

`transform()` (`evalml.pipelines.components.transformers.DelayedFeatureTransformer`), 791

`transform()` (`evalml.pipelines.components.transformers.DESTransformer`), 794

`transform()` (`evalml.pipelines.components.transformers.dimensionalities.LSA`), 645

`transform()` (`evalml.pipelines.components.transformers.dimensionalities.LSA`), 650

`transform()` (`evalml.pipelines.components.transformers.dimensionalities.LSA`), 652

`transform()` (`evalml.pipelines.components.transformers.dimensionalities.LSA`), 647

`transform()` (`evalml.pipelines.components.transformers.DropColumn`), 796

`transform()` (`evalml.pipelines.components.transformers.DropSubColumns`), 798

`transform()` (`evalml.pipelines.components.transformers.encoder.OneHotEncoder`), 656

`transform()` (`evalml.pipelines.components.transformers.encoder.OneHotEncoder`), 663

`transform()` (`evalml.pipelines.components.transformers.encoder.TargetEncoder`), 659

`transform()` (`evalml.pipelines.components.transformers.encoder.TargetEncoder`), 665

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 668

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 677

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 671

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 674

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 680

`transform()` (`evalml.pipelines.components.transformers.feature_selection.FeatureSelector`), 683

`transform()` (`evalml.pipelines.components.transformers.FeatureSelector`), 800

`transform()` (`evalml.pipelines.components.transformers.Imputer`), 802

`transform()` (`evalml.pipelines.components.transformers.imputers.Imputer`), 697

`transform()` (`evalml.pipelines.components.transformers.imputers.imputer`), 686

`transform()` (`evalml.pipelines.components.transformers.imputers.per_column`), 689

`transform()` (`evalml.pipelines.components.transformers.imputers.PerColumn`), 700

`transform()` (`evalml.pipelines.components.transformers.imputers.simple`), 691

`transform()` (`evalml.pipelines.components.transformers.imputers.Simple`), 702

`transform()` (`evalml.pipelines.components.transformers.imputers.target`), 694

`transform()` (`evalml.pipelines.components.transformers.imputers.Target`), 704

`transform()` (`evalml.pipelines.components.transformers.LinearDiscriminantAnalysis`), 805

`transform()` (`evalml.pipelines.components.transformers.LSA`), 807

`transform()` (`evalml.pipelines.components.transformers.LSA`), 810

`transform()` (`evalml.pipelines.components.transformers.LSA`), 812

`transform()` (`evalml.pipelines.components.transformers.LSA`), 814

`transform()` (`evalml.pipelines.components.transformers.PolynomialDet`), 817

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 707

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 728

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 710

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 731

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 733

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 712

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 735

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 715

`transform()` (`evalml.pipelines.components.transformers.preprocessing`), 737

U

[Undersampler](#) (class in [method](#)), 196
[evalml.pipelines.components](#)), 978
[Undersampler](#) (class in [method](#)), 218
[evalml.pipelines.components.transformers](#)),
845 [method](#)), 198
[Undersampler](#) (class in [method](#)), 219
[evalml.pipelines.components.transformers.samplers](#)),
767 [method](#)), 199
[Undersampler](#) (class in [method](#)), 220
[evalml.pipelines.components.transformers.samplers.undersampler](#)),
757 [method](#)), 220
[uniqueness_score\(\)](#) [method](#)), 200
[\(evalml.data_checks.uniqueness_data_check.UniquenessDataCheck](#)
[static method\)](#), 206
[uniqueness_score\(\)](#) [method](#)), 220
[\(evalml.data_checks.UniquenessDataCheck](#)
[static method\)](#), 225
[UniquenessDataCheck](#) (class in [method](#)), 221
[evalml.data_checks](#)), 224
[UniquenessDataCheck](#) (class in [method](#)), 202
[evalml.data_checks.uniqueness_data_check](#)),
206 [method](#)), 221
[update\(\)](#) ([evalml.automl.pipeline_search_plots.SearchIterationPlot](#)
[method](#)), 173 [method](#)), 204
[method](#)), 223
[method](#)), 205
[method](#)), 224
[method](#)), 207
[method](#)), 225
[method](#)), 208
[method](#)), 214
[method](#)), 193
[method](#)), 215
[method](#)), 194
[method](#)), 216
[method](#)), 216
[method](#)), 195
[method](#)), 217

V

[validate\(\)](#) ([evalml.data_checks.class_imbalance_data_check.ClassImbalanceDataCheck](#)
[method](#)), 184
[validate\(\)](#) ([evalml.data_checks.ClassImbalanceDataCheck](#)
[method](#)), 209
[validate\(\)](#) ([evalml.data_checks.data_check.DataCheck](#)
[method](#)), 186
[validate\(\)](#) ([evalml.data_checks.data_checks.DataChecks](#)
[method](#)), 192
[validate\(\)](#) ([evalml.data_checks.DataCheck](#) [method](#)),
211 [method](#)), 208
[validate\(\)](#) ([evalml.data_checks.DataChecks](#)
[method](#)), 214
[validate\(\)](#) ([evalml.data_checks.datetime_nan_data_check.DateTimeNaNDataCheck](#)
[method](#)), 193
[validate\(\)](#) ([evalml.data_checks.DateTimeNaNDataCheck](#)
[method](#)), 215
[validate\(\)](#) ([evalml.data_checks.default_data_checks.DefaultDataChecks](#)
[method](#)), 194
[validate\(\)](#) ([evalml.data_checks.DefaultDataChecks](#)
[method](#)), 216
[validate\(\)](#) ([evalml.data_checks.EmptyDataChecks](#)
[method](#)), 216
[validate\(\)](#) ([evalml.data_checks.highly_null_data_check.HighlyNullDataCheck](#)
[method](#)), 195
[validate\(\)](#) ([evalml.data_checks.HighlyNullDataCheck](#)
[method](#)), 217
[validate\(\)](#) ([evalml.data_checks.id_columns_data_check.IDColumnsDataCheck](#)
[method](#)), 196
[validate\(\)](#) ([evalml.data_checks.IDColumnsDataCheck](#)
[method](#)), 218
[validate\(\)](#) ([evalml.data_checks.invalid_targets_data_check.InvalidTargetsDataCheck](#)
[method](#)), 198
[validate\(\)](#) ([evalml.data_checks.InvalidTargetDataCheck](#)
[method](#)), 219
[validate\(\)](#) ([evalml.data_checks.multicollinearity_data_check.MulticollinearityDataCheck](#)
[method](#)), 199
[validate\(\)](#) ([evalml.data_checks.MulticollinearityDataCheck](#)
[method](#)), 220
[validate\(\)](#) ([evalml.data_checks.natural_language_nan_data_check.NaturalLanguageNaNDataCheck](#)
[method](#)), 220
[validate\(\)](#) ([evalml.data_checks.no_variance_data_check.NoVarianceDataCheck](#)
[method](#)), 201
[validate\(\)](#) ([evalml.data_checks.NoVarianceDataCheck](#)
[method](#)), 221
[validate\(\)](#) ([evalml.data_checks.outliers_data_check.OutliersDataCheck](#)
[method](#)), 202
[validate\(\)](#) ([evalml.data_checks.OutliersDataCheck](#)
[method](#)), 221
[validate\(\)](#) ([evalml.data_checks.sparsity_data_check.SparsityDataCheck](#)
[method](#)), 204
[validate\(\)](#) ([evalml.data_checks.SparsityDataCheck](#)
[method](#)), 223
[validate\(\)](#) ([evalml.data_checks.target_leakage_data_check.TargetLeakageDataCheck](#)
[method](#)), 205
[validate\(\)](#) ([evalml.data_checks.TargetLeakageDataCheck](#)
[method](#)), 224
[validate\(\)](#) ([evalml.data_checks.uniqueness_data_check.UniquenessDataCheck](#)
[method](#)), 207
[validate\(\)](#) ([evalml.data_checks.UniquenessDataCheck](#)
[method](#)), 225
[validate\(\)](#) ([evalml.data_checks.utils.EmptyDataChecks](#)
[method](#)), 208
[validate_inputs\(\)](#) ([evalml.objectives.AccuracyBinary](#) [method](#)),
352
[validate_inputs\(\)](#) ([evalml.objectives.AccuracyMulticlass](#) [method](#)),
354
[validate_inputs\(\)](#) ([evalml.objectives.AUC](#) [method](#)), 356
[validate_inputs\(\)](#) ([evalml.objectives.AUCMacro](#) [method](#)), 357
[validate_inputs\(\)](#) ([evalml.objectives.AUCMicro](#) [method](#)), 359
[validate_inputs\(\)](#) ([evalml.objectives.AUCWeighted](#) [method](#)),
360
[validate_inputs\(\)](#)

```

        (evalml.objectives.BalancedAccuracyBinary
         method), 363
validate_inputs()
    (evalml.objectives.BalancedAccuracyMulticlass
     method), 364
validate_inputs()
    (evalml.objectives.binary_classification_objective.BinaryClassificationObjective
     method), 267
validate_inputs()
    (evalml.objectives.BinaryClassificationObjective
     method), 367
validate_inputs()
    (evalml.objectives.cost_benefit_matrix.CostBenefitMatrix
     method), 270
validate_inputs()
    (evalml.objectives.CostBenefitMatrix
     method), 369
validate_inputs() (evalml.objectives.ExpVariance
                    method), 371
validate_inputs() (evalml.objectives.F1
                    method), 373
validate_inputs() (evalml.objectives.F1Macro
                    method), 375
validate_inputs() (evalml.objectives.F1Micro
                    method), 376
validate_inputs() (evalml.objectives.F1Weighted
                    method), 378
validate_inputs()
    (evalml.objectives.fraud_cost.FraudCost
     method), 273
validate_inputs() (evalml.objectives.FraudCost
                    method), 380
validate_inputs()
    (evalml.objectives.lead_scoring.LeadScoring
     method), 275
validate_inputs() (evalml.objectives.LeadScoring
                    method), 383
validate_inputs()
    (evalml.objectives.LogLossBinary
     method), 385
validate_inputs()
    (evalml.objectives.LogLossMulticlass
     method), 387
validate_inputs() (evalml.objectives.MAE
                    method), 388
validate_inputs() (evalml.objectives.MAPE
                    method), 390
validate_inputs() (evalml.objectives.MaxError
                    method), 392
validate_inputs() (evalml.objectives.MCCBinary
                    method), 394
validate_inputs()
    (evalml.objectives.MCCMulticlass
     method), 395
validate_inputs()
    (evalml.objectives.MeanSquaredLogError
     method), 397
validate_inputs() (evalml.objectives.MedianAE
                    method), 399
validate_inputs() (evalml.objectives.MSE
                    method), 400
validate_inputs()
    (evalml.objectives.multiclass_classification_objective.MulticlassClassificationObjective
     method), 278
validate_inputs()
    (evalml.objectives.MulticlassClassificationObjective
     method), 402
validate_inputs()
    (evalml.objectives.objective_base.ObjectiveBase
     method), 280
validate_inputs()
    (evalml.objectives.ObjectiveBase
     method), 404
validate_inputs() (evalml.objectives.Precision
                    method), 406
validate_inputs()
    (evalml.objectives.PrecisionMacro
     method), 408
validate_inputs()
    (evalml.objectives.PrecisionMicro
     method), 409
validate_inputs()
    (evalml.objectives.PrecisionWeighted
     method), 411
validate_inputs() (evalml.objectives.R2
                    method), 413
validate_inputs() (evalml.objectives.Recall
                    method), 415
validate_inputs()
    (evalml.objectives.RecallMacro
     method), 416
validate_inputs() (evalml.objectives.RecallMicro
                    method), 418
validate_inputs()
    (evalml.objectives.RecallWeighted
     method), 419
validate_inputs()
    (evalml.objectives.regression_objective.RegressionObjective
     method), 282
validate_inputs()
    (evalml.objectives.RegressionObjective
     method), 421
validate_inputs()
    (evalml.objectives.RootMeanSquaredError
     method), 423
validate_inputs()
    (evalml.objectives.RootMeanSquaredLogError
     method), 425

```

<code>validate_inputs()</code> (<code>evalml.objectives.sensitivity_low_alert.SensitivityLowAlert</code> <code>method</code>), 285	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MAPE</code> <code>method</code>), 316
<code>validate_inputs()</code> (<code>evalml.objectives.SensitivityLowAlert</code> <code>method</code>), 427	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MaxError</code> <code>method</code>), 318
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AccuracyBinary</code> <code>method</code>), 288	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MCCBinary</code> <code>method</code>), 320
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AccuracyMulticlass</code> <code>method</code>), 290	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MCCMulticlass</code> <code>method</code>), 321
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AUC</code> <code>method</code>), 292	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MeanSquaredLogError</code> <code>method</code>), 323
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AUCMacro</code> <code>method</code>), 294	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MedianAE</code> <code>method</code>), 324
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AUCMicro</code> <code>method</code>), 295	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MSE</code> <code>method</code>), 326
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.AUCWeighted</code> <code>method</code>), 297	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.Precision</code> <code>method</code>), 328
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.BalancedAccuracyBinary</code> <code>method</code>), 299	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.PrecisionMacro</code> <code>method</code>), 330
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.BalancedAccuracyMulticlass</code> <code>method</code>), 301	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.PrecisionMicro</code> <code>method</code>), 331
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.ExpVariance</code> <code>method</code>), 302	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.PrecisionWeighted</code> <code>method</code>), 333
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.F1</code> <code>method</code>), 304	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.R2</code> <code>method</code>), 334
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.F1Macro</code> <code>method</code>), 306	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.Recall</code> <code>method</code>), 337
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.F1Micro</code> <code>method</code>), 307	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.RecallMacro</code> <code>method</code>), 338
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.F1Weighted</code> <code>method</code>), 309	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.RecallMicro</code> <code>method</code>), 340
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.LogLossBinary</code> <code>method</code>), 311	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.RecallWeighted</code> <code>method</code>), 341
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.LogLossMulticlass</code> <code>method</code>), 313	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.RootMeanSquaredError</code> <code>method</code>), 343
<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.MAE</code> <code>method</code>), 314	<code>validate_inputs()</code> (<code>evalml.objectives.standard_metrics.RootMeanSquaredLogError</code> <code>method</code>), 345

[validate_inputs\(\)](#) ([evalml.objectives.time_series_regression_objective.TimeSeriesRegressionObjective](#) (class in [evalml.pipelines.components.estimators](#)), method), 347
[value\(\)](#) ([evalml.data_checks.data_check_action_code.DataCheckActionCode](#) (class in [evalml.pipelines.components.estimators.classifiers](#)), method), 187
[value\(\)](#) ([evalml.data_checks.data_check_message_code.DataCheckMessageCode](#) (class in [evalml.pipelines.components.estimators.classifiers.xgboost_classifier](#)), method), 191
[value\(\)](#) ([evalml.data_checks.data_check_message_type.DataCheckMessageType](#) (class in [evalml.pipelines.components.estimators.classifiers.xgboost_classifier](#)), method), 191
[value\(\)](#) ([evalml.data_checks.DataCheckActionCode](#) (class in [evalml.pipelines](#)), method), 211
[value\(\)](#) ([evalml.data_checks.DataCheckMessageCode](#) (class in [evalml.pipelines.components](#)), method), 214
[value\(\)](#) ([evalml.data_checks.DataCheckMessageType](#) (class in [evalml.pipelines.components.estimators](#)), method), 214
[value\(\)](#) ([evalml.model_family.model_family.ModelFamily](#) (class in [evalml.pipelines.components.estimators.regressors](#)), method), 232
[value\(\)](#) ([evalml.model_family.ModelFamily](#) (class in [evalml.pipelines.components.estimators.regressors](#)), method), 234
[value\(\)](#) ([evalml.model_understanding.prediction_explanations.explanations.Explanations](#) (class in [evalml.pipelines.components.estimators.regressors.xgboost_regressor](#)), method), 238
[value\(\)](#) ([evalml.problem_types.problem_types.ProblemTypes](#) (class in [evalml.pipelines.components](#)), method), 1172
[value\(\)](#) ([evalml.problem_types.ProblemTypes](#) (class in [evalml.pipelines.components](#)), method), 1177
[visualize_decision_tree\(\)](#) (in module [evalml.model_understanding.graphs](#)), 251

W

[warning\(\)](#) ([evalml.automl.engine.engine_base.JobLogger](#) (class in [evalml.pipelines.components](#)), method), 157
[warning_not_unique_enough](#) (in module [evalml.data_checks.uniqueness_data_check](#)), 207
[warning_too_unique](#) (in module [evalml.data_checks.sparsity_data_check](#)), 204
[warning_too_unique](#) (in module [evalml.data_checks.uniqueness_data_check](#)), 207
[WrappedSKClassifier](#) (class in [evalml.pipelines.components.utils](#)), 853
[WrappedSKRegressor](#) (class in [evalml.pipelines.components.utils](#)), 854
[write_to_logger\(\)](#) ([evalml.automl.engine.engine_base.JobLogger](#) (class in [evalml.pipelines.components](#)), method), 157

X

[XGBoostClassifier](#) (class in [evalml.pipelines](#)), 1155
[XGBoostClassifier](#) (class in [evalml.pipelines.components](#)), 981