toad

Release 0.0.57

Nov 06, 2019

Contents

CHAPTER 1

Installation

via pip

```
pip install toad
```

via source code

```
python setup.py install
```

CHAPTER 2

Tutorial

A basic tutorial is provided.

CHAPTER 3

Contents

3.1 toad package

3.2 Submodules

3.2.1 toad.detector module

Command line tools for detecting csv data

Team: ESC

Examples

python detector.py -i xxx.csv -o report.csv

toad.detector.**countBlank**(*series, blanks=[None]*)

Count number and percentage of blank values in series

Parameters

- **series** (*Series*) – data series
- **blanks** (*list*) – list of blank values

Returns number of blanks str: the percentage of blank values

Return type number

toad.detector.**detect**(*dataframe*)

Detect data

Parameters **dataframe** (*DataFrame*) – data that will be detected

Returns report of detecting

Return type DataFrame

toad.detector.**getDescribe**(*series, percentiles=[0.25, 0.5, 0.75]*)
Get describe of series

Parameters

- **series** (*Series*) – data series
- **percentiles** – the percentiles to include in the output

Returns the describe of data include mean, std, min, max and percentiles

Return type Series

toad.detector.**getTopValues**(*series*, *top=5*, *reverse=False*)
Get top/bottom n values

Parameters

- **series** (*Series*) – data series
- **top** (*number*) – number of top/bottom n values
- **reverse** (*bool*) – it will return bottom n values if True is given

Returns Series of top/bottom n values and percentage. [‘value:percent’, None]

Return type Series

toad.detector.**isNumeric**(*series*)
Check if the series’s type is numeric

Parameters **series** (*Series*) – data series

Returns bool

3.2.2 toad.merge module

toad.merge.**ChiMerge**()
Chi-Merge

Parameters

- **feature** (*array-like*) – feature to be merged
- **target** (*array-like*) – a array of target classes
- **n_bins** (*int*) – n bins will be merged into
- **min_samples** (*number*) – min sample in each group, if float, it will be the percentage of samples
- **min_threshold** (*number*) – min threshold of chi-square

Returns array of split points

Return type array

toad.merge.**DTMerge**()
Merge continue

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) – target will be used to fit decision tree
- **nan** (*number*) – value will be used to fill nan

- **n_bins** (*int*) – n groups that will be merged into
- **min_samples** (*int*) – min number of samples in each leaf nodes

Returns array of split points

Return type array

toad.merge.**KMeansMerge**()
Merge by KMeans

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) – target will be used to fit kmeans model
- **nan** (*number*) – value will be used to fill nan
- **n_bins** (*int*) – n groups that will be merged into
- **random_state** (*int*) – random state will be used for kmeans model

Returns split points of feature

Return type array

toad.merge.**QuantileMerge**()
Merge by quantile

Parameters

- **feature** (*array-like*) –
- **nan** (*number*) – value will be used to fill nan
- **n_bins** (*int*) – n groups that will be merged into
- **q** (*array-like*) – list of percentage split points

Returns split points of feature

Return type array

toad.merge.**StepMerge**()
Merge by step

Parameters

- **feature** (*array-like*) –
- **nan** (*number*) – value will be used to fill nan
- **n_bins** (*int*) – n groups that will be merged into
- **clip_v** (*number | tuple*) – min/max value of clipping
- **clip_std** (*number | tuple*) – min/max std of clipping
- **clip_q** (*number | tuple*) – min/max quantile of clipping

Returns split points of feature

Return type array

toad.merge.**merge**()
merge feature into groups

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) –
- **method** (*str*) – 'dt', 'chi', 'quantile', 'step', 'kmeans' - the strategy to be used to merge feature
- **return_splits** (*bool*) – if needs to return splits
- **n_bins** (*int*) – n groups that will be merged into

Returns a array of merged label with the same size of feature array: list of split points

Return type array

3.2.3 toad.metrics module

toad.metrics.**AIC**(*y_pred*, *y*, *k*, *lf=None*)
Akaike Information Criterion

Parameters

- **y_pred** (*array-like*) –
- **y** (*array-like*) –
- **k** (*int*) – number of featuers
- **lf** (*float*) – result of likelihood function

toad.metrics.**AUC**(*score*, *target*)
AUC Score

Parameters

- **score** (*array-like*) – list of score or probability that the model predict
- **target** (*array-like*) – list of real target

Returns auc score

Return type float

toad.metrics.**BIC**(*y_pred*, *y*, *k*, *lf=None*)
Bayesian Information Criterion

Parameters

- **y_pred** (*array-like*) –
- **y** (*array-like*) –
- **k** (*int*) – number of featuers
- **lf** (*float*) – result of likelihood function

toad.metrics.**F1**(*score*, *target*, *split='best'*, *return_split=False*)
calculate f1 value

Parameters

- **score** (*array-like*) –
- **target** (*array-like*) –

Returns best f1 score float: best spliter

Return type float

toad.metrics.**KS**(*score*, *target*)
calculate ks value

Parameters

- **score** (*array-like*) – list of score or probability that the model predict
- **target** (*array-like*) – list of real target

Returns the max KS value

Return type float

toad.metrics.**KS_bucket**(*score*, *target*, *bucket=10*, *method='quantile'*, ***kwargs*)
calculate ks value by bucket

Parameters

- **score** (*array-like*) – list of score or probability that the model predict
- **target** (*array-like*) – list of real target
- **bucket** (*int*) – n groups that will bin into
- **method** (*str*) – method to bin score. *quantile* (default), *step*

Returns DataFrame

toad.metrics.**KS_by_col**(*df*, *by='feature'*, *score='score'*, *target='target'*)

toad.metrics.**MSE**(*y_pred*, *y*)
mean of squares due to error

toad.metrics.**PSI**(*test*, *base*, *combiner=None*, *return_frame=False*)
calculate PSI

Parameters

- **test** (*array-like*) – data to test PSI
- **base** (*array-like*) – base data for calculate PSI
- **combiner** (*Combiner|list|dict*) – combiner to combine data
- **return_frame** (*bool*) – if need to return frame of proportion

Returns float|Series

toad.metrics.**SSE**(*y_pred*, *y*)
sum of squares due to error

3.2.4 toad.plot module

toad.plot.**badrate_plot**(*frame*, *x=None*, *target='target'*, *by=None*, *freq=None*, *format=None*, *return_counts=False*, *return_proportion=False*, *return_frame=False*)
plot for badrate

Parameters

- **frame** (*DataFrame*) –
- **x** (*str*) – column in frame that will be used as x axis
- **target** (*str*) – target column in frame
- **by** (*str*) – column in frame that will be calculated badrate by it

- **freq** (*str*) – offset aliases string by pandas http://pandas.pydata.org/pandas-docs/stable/timeseries.html#offset-aliases
- **format** (*str*) – format string for time
- **return_counts** (*bool*) – if need return counts plot
- **return_frame** (*bool*) – if need return frame

Returns badrate plot Axes: counts plot Axes: proportion plot Dataframe: grouping detail data

Return type Axes

toad.plot.**bin_plot**(*frame*, *x=None*, *target='target'*, *iv=True*)
plot for bins

toad.plot.**corr_plot**(*frame*, *figure_size=(20, 15)*)
plot for correlation

Parameters **frame** (*DataFrame*) – frame to draw plot

Returns Axes

toad.plot.**proportion_plot**(*x=None*, *keys=None*)
plot for proportion

Parameters

- **x** (*Series|list*) – series or list of series data for plot
- **keys** (*str|list*) – keys for each data

Returns Axes

toad.plot.**roc_plot**(*score*, *target*)
plot for roc

Parameters

- **score** (*array-like*) – predicted score
- **target** (*array-like*) – true target

Returns Axes

3.2.5 toad.scorecard module

class toad.scorecard.**ScoreCard**(*pdo=60*, *rate=2*, *base_odds=35*, *base_score=750*, ***kwargs*)
Bases: sklearn.base.BaseEstimator

bin_to_score(*bins*, *return_sub=False*)
predict score from bins

combine(*X*)

export(*to_frame=False*, *to_json=None*, *to_csv=None*, *decimal=2*)
generate a scorecard object

Parameters

- **to_frame** (*bool*) – return DataFrame of card
- **to_json** (*str|IOBase*) – io to write json file
- **to_csv** (*filepath|IOBase*) – file to write csv

Returns dict

fit(*X*, *y*, *combiner=None*, *transer=None*, *model=None*)

Parameters

- **X** (*2D array-like*) –
- **Y** (*array-like*) –

generate_card(*card=None*, *combiner={}*, *transer=None*, *model=None*)

Parameters

- **card** (*dict | str | IOBase*) – dict of card or io to read json
- **combiner** (*toad.Combiner*) –
- **transer** (*toad.WOETransformer*) –
- **model** (*LogisticRegression*) –

generate_map(*transer*, *model*)
calculate score map by woe

predict(*X*, ***kwargs*)
predict score :param X: X to predict :type X: 2D array-like :param return_sub: if need to return sub score, default *False* :type return_sub: bool

Returns predicted score DataFrame: sub score for each feature

Return type array-like

proba_to_score(*prob*)
covert probability to score

set_card(*card*)
set card dict

set_combiner(*combiner*)
set combiner

set_model(*model*)
set logistic regression model

set_score(*map*)
set score map by dict

testing_frame(***kwargs*)
get testing frame with score

Returns testing frame with score

Return type DataFrame

woe_to_score(*woe*, *weight=None*)
calculate score by woe

3.2.6 toad.selection module

class toad.selection.**StatsModel**(*estimator='ols'*, *criterion='aic'*, *intercept=False*)
Bases: object

get_criterion(*pre*, *y*, *k*)

get_estimator(*name*)

likelihood(*pre*, *y*, *k*)

p_value(*t*, *n*)

stats(*X*, *y*)

t_value(*pre*, *y*, *X*, *coef*)

toad.selection.**drop_corr**(*frame*, *target=None*, *threshold=0.7*, *by='IV'*, *return_drop=False*, *exclude=None*)

drop columns by correlation

Parameters

- **frame** (*DataFrame*) – dataframe that will be used
- **target** (*str*) – target name in dataframe
- **threshold** (*float*) – drop features that has the smallest weight in each groups whose correlation is greater than threshold
- **by** (*array-like*) – weight of features that will be used to drop the features
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped

Return type DataFrame

toad.selection.**drop_empty**(*frame*, *threshold=0.9*, *nan=None*, *return_drop=False*, *exclude=None*)

drop columns by empty

Parameters

- **frame** (*DataFrame*) – dataframe that will be used
- **threshold** (*number*) – drop the features whose empty num is greater than threshold. if threshold is float, it will be use as percentage
- **nan** (*any*) – values will be look like empty
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped

Return type DataFrame

toad.selection.**drop_iv**(*frame*, *target='target'*, *threshold=0.02*, *return_drop=False*, *return_iv=False*, *exclude=None*)

drop columns by IV

Parameters

- **frame** (*DataFrame*) – dataframe that will be used
- **target** (*str*) – target name in dataframe
- **threshold** (*float*) – drop the features whose IV is less than threshold
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **return_iv** (*bool*) – if need to return features' IV
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped Series: list of features' IV

Return type DataFrame

toad.selection.**drop_var**(*frame*, *threshold=0*, *return_drop=False*, *exclude=None*)
drop columns by variance

Parameters

- **frame** (*DataFrame*) – dataframe that will be used
- **threshold** (*float*) – drop features whose variance is less than threshold
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped

Return type DataFrame

toad.selection.**drop_vif**(*frame*, *threshold=3*, *return_drop=False*, *exclude=None*)
variance inflation factor

Parameters

- **frame** (*DataFrame*) –
- **threshold** (*float*) – drop features until all vif is less than threshold
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped

Return type DataFrame

toad.selection.**select**(*frame*, *target='target'*, *empty=0.9*, *iv=0.02*, *corr=0.7*, *return_drop=False*, *exclude=None*)
select features by rate of empty, iv and correlation

Parameters

- **frame** (*DataFrame*) –
- **target** (*str*) – target's name in dataframe
- **empty** (*number*) – drop the features which empty num is greater than threshold. if threshold is float, it will be use as percentage
- **iv** (*float*) – drop the features whose IV is less than threshold
- **corr** (*float*) – drop features that has the smallest IV in each groups which correlation is greater than threshold
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature name that will not be dropped

Returns selected dataframe dict: list of dropped feature names in each step

Return type DataFrame

toad.selection.**stepwise**(*frame*, *target='target'*, *estimator='ols'*, *direction='both'*, *criterion='aic'*, *p_enter=0.01*, *p_remove=0.01*, *p_value_enter=0.2*, *intercept=False*, *max_iter=None*, *return_drop=False*, *exclude=None*)
stepwise to select features

Parameters

- **frame** (*DataFrame*) – dataframe that will be use to select
- **target** (*str*) – target name in frame
- **estimator** (*str*) – model to use for stats
- **direction** (*str*) – direction of stepwise, support 'forward', 'backward' and 'both', suggest 'both'
- **criterion** (*str*) – criterion to statistic model, support 'aic', 'bic'
- **p_enter** (*float*) – threshold that will be used in 'forward' and 'both' to keep features
- **p_remove** (*float*) – threshold that will be used in 'backward' to remove features
- **intercept** (*bool*) – if have intercept
- **p_value_enter** (*float*) – threshold that will be used in 'both' to remove features
- **max_iter** (*int*) – maximum number of iterate
- **return_drop** (*bool*) – if need to return features' name who has been dropped
- **exclude** (*array-like*) – list of feature names that will not be dropped

Returns selected dataframe array: list of feature names that has been dropped

Return type DataFrame

3.2.7 toad.stats module

toad.stats.**IV**(*feature*, *target*, ***kwargs*)
get the IV of a feature

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) –
- **n_bins** (*int*) – n groups that the feature will bin into
- **method** (*str*) – the strategy to be used to merge feature, default is 'dt'
- **()** (***kwargs*) – other options for merge function

toad.stats.**VIF**(*frame*)
calculate vif

Parameters **frame** (*ndarray|DataFrame*) –

Returns Series

toad.stats.**WOE**(*y_prob*, *n_prob*)
get WOE of a group

Parameters

- **y_prob** – the probability of grouped y in total y
- **n_prob** – the probability of grouped n in total n

Returns woe value

Return type number

toad.stats.**badrate**(*target*)
calculate badrate

Parameters **target** (*array-like*) – target array which *1* is bad

Returns float

toad.stats.**column_quality**(*feature*, *target*, *name='feature'*, *iv_only=False*, ***kwargs*)
calculate quality of a feature

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) –
- **name** (*str*) – feature's name that will be setted in the returned Series
- **iv_only** (*bool*) – if only calculate IV

Returns a list of quality with the feature's name

Return type Series

toad.stats.**entropy**(*target*)
get infomation entropy of a feature

Parameters **target** (*array-like*) –

Returns information entropy

Return type number

toad.stats.**entropy_cond**(*feature*, *target*)
get conditional entropy of a feature

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) –

Returns conditional information entropy. If feature is continuous, it will return the best entropy when the feature bins into two groups

Return type number

toad.stats.**gini**(*target*)
get gini index of a feature

Parameters **target** (*array-like*) – list of target that will be calculate gini

Returns gini value

Return type number

toad.stats.**gini_cond**(*feature*, *target*)
get conditional gini index of a feature

Parameters

- **feature** (*array-like*) –
- **target** (*array-like*) –

Returns conditional gini value. If feature is continuous, it will return the best gini value when the feature bins into two groups

Return type number

toad.stats.**probability**(*target*, *mask=None*)
get probability of target by mask

toad.stats.**quality**(*dataframe*, *target='target'*, *iv_only=False*, ***kwargs*)
get quality of features in data

Parameters

- **dataframe** (*DataFrame*) – dataframe that will be calculate quality
- **target** (*str*) – the target's name in dataframe
- **iv_only** (*bool*) – if only calculate IV

Returns quality of features with the features' name as row name

Return type DataFrame

3.2.8 toad.transform module

class toad.transform.**Combiner**
Bases: sklearn.base.TransformerMixin

Combiner for merge data

dtypes
get the dtypes which is combiner used

Returns (str|dict)

export(*format=False*)
export combine rules for score card

Parameters

- **format** (*bool*) – if True, bins will be replace with string label for values
- **to_json** (*str|IOBase*) – io to write json file

Returns dict

fit(*X*, *y=None*, ***kwargs*)
fit combiner

Parameters

- **X** (*DataFrame|array-like*) – features to be combined
- **y** (*str|array-like*) – target data or name of target in *X*
- **method** (*str*) – the strategy to be used to merge *X*, same as *.merge*, default is *chi*
- **n_bins** (*int*) – counts of bins will be combined

Returns self

set_rules(*map*, *reset=False*)
set rules for combiner

Parameters

- **map** (*dict|array-like*) – map of splits
- **reset** (*bool*) – if need to reset combiner

Returns self

transform(*X*, ***kwargs*)
transform X by combiner

Parameters

- **X** (*DataFrame|array-like*) – features to be transformed
- **labels** (*bool*) – if need to use labels for resulting bins, *False* by default

Returns array-like

class toad.transform.**WOETransformer**
Bases: sklearn.base.TransformerMixin

WOE transformer

export()

fit(*X*, *y*, ***kwargs*)
fit WOE transformer

Parameters

- **X** (*DataFrame|array-like*) –
- **y** (*str|array-like*) –
- **select_dtypes** (*str|numpy.dtypes*) – *'object'*, *'number'* etc. only selected dtypes will be transform,

transform(*X*, ***kwargs*)
transform woe

Parameters

- **X** (*DataFrame|array-like*) –
- **default** (*str*) – 'min'(default), 'max' - the strategy to be used for unknown group

Returns array-like

toad.transform.**support_exclude**(*fn*)

toad.transform.**support_save_to_json**(*fn*)

toad.transform.**support_select_dtypes**(*fn*)

3.2.9 toad.utils module

class toad.utils.**Parallel**
Bases: object

apply(*func*, *args=()*, *kwargs={}*)

join()

toad.utils.**bin_by_splits**(*feature*, *splits*)
Bin feature by split points

toad.utils.**bin_to_number**(*reg=None*)

Returns func(string) -> number

Return type function

toad.utils.**clip**(*series*, *value=None*, *std=None*, *quantile=None*)
clip series

Parameters

- **series** (*array-like*) – series need to be clipped
- **value** (*number | tuple*) – min/max value of clipping
- **std** (*number | tuple*) – min/max std of clipping
- **quantile** (*number | tuple*) – min/max quantile of clipping

toad.utils.**diff_time**(*base*, *target*, *format=None*, *time='day'*)

toad.utils.**diff_time_frame**(*base*, *frame*, *format=None*)

toad.utils.**feature_splits**(*feature*, *target*)
find posibility spilt points

toad.utils.**fillna**(*feature*, *by=-1*)

toad.utils.**generate_str**(*size=6*, *chars='ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789'*)

toad.utils.**generate_target**(*size*, *rate=0.5*, *weight=None*, *reverse=False*)
generate target for reject inference

Parameters

- **size** (*int*) – size of target
- **rate** (*float*) – rate of '1' in target
- **weight** (*array-like*) – weight of '1' to generate target
- **reverse** (*bool*) – if need reverse weight

Returns array

toad.utils.**get_dummies**(*dataframe*, *exclude=None*, *binary_drop=False*, ***kwargs*)
get dummies

toad.utils.**has_nan**(*arr*)

toad.utils.**inter_feature**(*feature*, *splits*)

toad.utils.**is_continuous**(*series*)

toad.utils.**iter_df**(*dataframe*, *feature*, *target*, *splits*)
iterate dataframe by split points

Returns iterator (df, splitter)

toad.utils.**np_count**(*arr*, *value*, *default=None*)

toad.utils.**np_unique**(*arr*, ***kwargs*)

toad.utils.**read_json**(*file*)
read json file

toad.utils.**save_json**(*contents*, *file*, *indent=4*)
save json file

Parameters

- **contents** (*dict*) – contents to save
- **file** (*str|IOBase*) – file to save

toad.utils.**split_target**(*frame*, *target*)

toad.utils.**support_dataframe**(*require_target=True*)
decorator for supporting dataframe

toad.utils.**to_ndarray**(*s*, *dtype=None*)

toad.utils.**unpack_tuple**(*x*)

3.3 Module contents

CHAPTER 4

Indices and tables

- genindex
- modindex
- search

Python Module Index

t

Index

J

K

L

M

N

P

Q

R

S

T

U

V

W