Package: TangledFeatures (via r-universe)

August 28, 2024

Type Package

Title Feature Selection in Highly Correlated Spaces

Version 0.1.1

Description Feature selection algorithm that extracts features in highly correlated spaces. The extracted features are meant to be fed into simple explainable models such as linear or logistic regressions. The package is useful in the field of explainable modelling as a way to understand variable behavior.

License MIT + file LICENSE

URL https://allen-1242.github.io/TangledFeatures/

Depends R (>= 2.10)

Imports correlation, data.table, dplyr, fastDummies, ggplot2, igraph, janitor, Matrix, methods, purrr, ranger, broom, broom.mixed, caret, jtools, randomForest, glmnet

Suggests knitr, R.rsp, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Repository https://allen-1242.r-universe.dev

RemoteUrl https://github.com/allen-1242/tangledfeatures

RemoteRef HEAD

RemoteSha 4fee69922fb905584365e35d3033976c38288432

Contents

`Advertisement` *Advertisement dataset*

Description

Advertisement dataset

`DataCleaning` *Automatic Data Cleaning*

Description

Automatic Data Cleaning

Usage

```
DataCleaning(Data, Y_var)
```

Arguments

`Data`	The imported Data Frame
`Y_var`	The X variable

Value

The cleaned data.

Examples

```
DataCleaning(Data = TangledFeatures::Housing_Prices_dataset,  Y_var = 'SalePrice')
```

`GeneralCor` *Generalized Correlation function*

Description

Generalized Correlation function

Usage

```
GeneralCor(df, cor1 = "pearson", cor2 = "polychoric", cor3 = "spearman")
```

Arguments

`df`	The imported Data Frame
`cor1`	The correlation metric between two continuous features. Defaults to pearson
`cor2`	The correlation metric between one categorical feature and one cont feature. Defaults to biserial
`cor3`	The correlation metric between two categorical features. Defaults to Cramers-V

Value

Returns a correlation matrix containing the correlation values between the features

Examples

```
GeneralCor(df = TangledFeatures::Advertisement)
```

`Housing_Prices_dataset`
Housing prices dataset

Description

Housing prices dataset

`TangledFeatures` *The main TangledFeatures function*

Description

The main TangledFeatures function

Usage

```
TangledFeatures(
  Data,
  Y_var,
  Focus_variables = list(),
  corr_cutoff = 0.85,
  RF_coverage = 0.95,
  plot = FALSE,
  fast_calculation = FALSE,
  cor1 = "pearson",
  cor2 = "polychoric",
  cor3 = "spearman"
)
```

Arguments

`Data`	The imported Data Frame
`Y_var`	The dependent variable
`Focus_variables`	The list of variables that you wish to give a certain bias to in the correlation matrix
`corr_cutoff`	The correlation cutoff variable. Defaults to 0.8
`RF_coverage`	The Random Forest coverage of explainable. Defaults to 95 percent
`plot`	Return if plotting is to be done. Binary True or False
`fast_calculation`	Returns variable list without many Random Forest iterations by simply picking a variable from a correlated group
`cor1`	The correlation metric between two continuous features. Defaults to pearson correlation
`cor2`	The correlation metric between one categorical feature and one continuous feature. Defaults to bi serial correlation correlation
`cor3`	The correlation metric between two categorical features. Defaults to Cramer's V.

Value

Returns a list of variables that are ready for future modelling, along with other metrics

Examples

```
TangledFeatures(Data = TangledFeatures::Advertisement, Y_var = 'Sales')
```

Index