Package ‘CollapseLevels’

October 12, 2022

Type Package

Title Collapses Levels, Computes Information Value and WoE

Version 0.3.0

Author Krishanu Mukherjee

Maintainer Krishanu Mukherjee <toton1181@gmail.com>

Description Contains functions to help in selecting and exploring features (or variables) in binary classification problems.
Provides functions to compute and display information value and weight of evidence (WoE) of the variables , and to convert numeric variables to categorical variables by binning.
Functions are also provided to determine which levels (or categories) of a categorical variable can be collapsed (or combined) based on their response rates.
The functions provided only work for binary classification problems.

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Imports dplyr,lazyeval, ggplot2

Depends magrittr

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-06-04 13:20:02 UTC

R topics documented:

`displayIV` *displayIV*

Description

This function displays the Information Values of the levels of an attribute.

Usage

```
displayIV(dset, col = "xyz", resp = "y", adjFactor = 0.5, bins = 10)
```

Arguments

`dset`	The data frame containing the data set
`col`	A character respresenting the name of the attribute . The attribute can either be numeric or categorical
`resp`	A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values
`adjFactor`	A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute
`bins`	A number denoting the number of bins.Default value is 10

Examples

```
# Load the German_Credit data set supplied with this package


data("German_Credit")

displayIV(German_Credit,col="Credit_History",resp="Good_Bad")
```

`displayResponseRatebyLevels` *displayResponseRatebyLevels*

Description

This function displays the response percents of the levels of an attribute.

Usage

```
displayResponseRatebyLevels(
  dset,
  col = "job",
  resp = "Good_Bad",
  bins = 10,
  adjFactor = 0.5
)
```

Arguments

`dset` The data frame containing the data set

`col` A character respresenting the name of the attribute . The attribute can either be numeric or categorical

`resp` A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values

`bins` A number denoting the number of bins.Default value is 10

`adjFactor` A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute

Examples

```
# Load the German_Credit data set supplied with this package

data("German_Credit")

displayResponseRatebyLevels(German_Credit,col="Credit_History",resp="Good_Bad")
```

displayWOE *displayWOE*

Description

This function displays the Weight of Evidence of the levels of an attribute.

Usage

```
displayWOE(dset, col = "xyz", resp = "y", adjFactor = 0.5, bins = 10)
```

Arguments

`dset`	The data frame containing the data set
`col`	A character respresenting the name of the attribute . The attribute can either be numeric or categorical
`resp`	A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values
`adjFactor`	A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute
`bins`	A number denoting the number of bins.Default value is 10

Examples

```
# Load the German_Credit data set supplied with this package

data("German_Credit")

displayWOE(German_Credit,col="Credit_History",resp="Good_Bad")
```

German_Credit *German Credit data set*

Description

This data set classifies customers as "Good" or "Bad" as per their credit risks.This data set was contributed by Professor Dr. Hans Hofmann,and can be downloaded from the UCI Machine Learning Repository.

Usage

```
data("German_Credit")
```

Format

A data frame with 1000 observations on the following 21 variables.

`Account_Balance` a factor with levels `A11` `A12` `A13` `A14`

`Duration` a numeric vector

`Credit_History` a factor with levels `A30` `A31` `A32` `A33` `A34`

`Purpose` a factor with levels `A40` `A41` `A410` `A42` `A43` `A44` `A45` `A46` `A48` `A49`

`Credit_Amount` a numeric vector

`Saving_Accounts_Bonds` a factor with levels `A61` `A62` `A63` `A64` `A65`

`Current_Employment_Length` a factor with levels `A71` `A72` `A73` `A74` `A75`

`Installment_Rate` a numeric vector

`MaritalStatusnGender` a factor with levels `A91` `A92` `A93` `A94`

`Guarantors` a factor with levels `A101` `A102` `A103`

‘`Duration in Current Address`’ a numeric vector

`Valuable_Asset` a factor with levels `A121` `A122` `A123` `A124`

`Age` a numeric vector

`Other_Credit` a factor with levels `A141` `A142` `A143`

`Housing` a factor with levels `A151` `A152` `A153`

`Existing_Credits` a numeric vector

`Job` a factor with levels `A171` `A172` `A173` `A174`

`Dependents` a numeric vector

`Telephone` a factor with levels `A191` `A192`

`ForeignWorker` a factor with levels `A201` `A202`

`Good_Bad` a numeric vector

Source

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Examples

```
data(German_Credit)
str(German_Credit)
```

IVCalc *IVCalc*

Description

This function displays the Information Values by the levels of an attribute This information is displayed for all attributes in the data set

Usage

```
IVCalc(dset, resp = "y", bins = 10, adjFactor = 0.5)
```

Arguments

`dset` The data frame containing the data set

`resp` A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values

`bins` A number denoting the number of bins.Default value is 10

`adjFactor` A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute

Value

A list containing the tables of Information Values by levels for every attribute

Examples

```
# Load the German_Credit data set supplied with this package

data("German_Credit")

l<-list()

# Call the function as follows

l<-IVCalc(German_Credit,resp="Good_Bad",bins=10)

# Information Value for  the attribute Account_Balance in the German_Credit data

l$Account_Balance
```

IVCalc2 *IVCalc2*

Description

This function displays the Information Values of all the attributes in the data set

Usage

```
IVCalc2(dset, resp = "y", bins = 10, adjFactor = 0.5)
```

Arguments

`dset`	The data frame containing the data set
`resp`	A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values
`bins`	A number denoting the number of bins.Default value is 10
`adjFactor`	A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute

Value

A data frame containing the Information Values for every attribute

Examples

```
# Load the German_Credit data set supplied with this package

data("German_Credit")

d<-data.frame()

# Call the function as follows

d<-IVCalc2(German_Credit,resp="Good_Bad",bins=10)

# Information Value for all the attributes in the German_Credit data

d
```

levelsCollapser *levelsCollapser*

Description

This function displays the response rates by the levels of an attribute Levels with similar response rates may be combined

Usage

```
levelsCollapser(dset, resp = "y", bins = 10)
```

Arguments

dset	The data frame containing the data set
resp	A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values
bins	A number denoting the number of bins.Default value is 10

Value

A list containing the tables of response rate by levels for every attribute

Examples

```

# Load the German_Credit data set supplied with this package

data("German_Credit")

# Create an empty list

l<-list()

# Call the function as follows

l<-levelsCollapser(German_Credit,resp="Good_Bad",bins=10)

# response rate by levels of the Account_Balance in the German_Credit data

l$Account_Balance

# Collapse levels with similar response percentages.
```

numericToCategorical *numericToCategorical*

Description

This function categorizes a numerical variable by binning

Usage

```
numericToCategorical(dset, col = "job", resp = "y", bins = 10, adjFactor = 0.5)
```

Arguments

dset	The data frame containing the data set
col	A character respresenting the name of the numeric attribute which we want to categorize
resp	A character respresenting the name of the binary outcome variable The binary outcome variable may be a factor with two levels or an integer (or numeric) with two unique values
bins	A number denoting the number of bins.Default value is 10
adjFactor	A number or a decimal denoting what is to be added to the number of responses (binary outcome variable is 1) or to the number of non responses (binary outcome variable is 0) if either is zero for any level of the attribute

Value

A list containing the categorized attribute,a table of Information Values for the levels of the categorized attribute,the Information Value for the entire attribute,a table showing the response rates of the levels of the categorized attribute

Examples

```
# Load the German_Credit data set supplied with this package

data("German_Credit")

# Create an empty list

l<-list()

# Call the function as follows.
#This will categorize the numeric variable Duration in the German_Credit dataset.

l<-numericToCategorical(German_Credit,col="Duration",resp="Good_Bad")
```

```
# To view the categorized variable

 l$categoricalVariable

 # To view the IV table of the levels of the categorized variable

 l$IVTable

 # To view the total IV value of the  categorized variable

 l$IV

 # To view the response rates of the levels of the categorized variable

 l$collapseLevels
```

Index