Package 'tree.bins'

June 14, 2018

Type Package

Title Recategorization of Factor Variables by Decision Tree Leaves

Version 0.1.1

Date 2018-06-13

Maintainer Piro Polo <piropolo98@gmail.com>

Description

Provides users the ability to categorize categorical variables dependent on a response variable. It creates a decision tree by using one of the categorical variables (class factor) and the selected response variable.
The decision tree is created from the rpart() function from the 'rpart' package. The rules from the leaves of
the decision tree are extracted, and used to recategorize the appropriate categorical variable (predictor). This
step is performed for each of the categorical variables that is fed into the data component of the function. Only variables containing more than 2 factor levels will be considered in the function. The final output generates a data set containing the recategorized variables or a list containing a mapping table
for each of the candidate variables. For more details see T. Hastie et al (2009, ISBN: 978-0-387-84857-0).

License GPL-2

Encoding UTF-8

LazyData TRUE

Depends R (>= 3.4.0)

Imports dplyr (>= 0.7.4), rpart (>= 4.1-11), rpart.utils (>= 0.5), data.table (>= 1.10.4-3)

RoxygenNote 6.0.1

Suggests knitr, rmarkdown, testthat, rpart.plot, ggplot2, ggthemes

VignetteBuilder knitr

NeedsCompilation no

Author Piro Polo [aut, cre]

Repository CRAN

Date/Publication 2018-06-14 05:33:53 UTC

R topics documented:

`AmesImp` *A Subset of the Ames Data Set with Imputed Values*

Description

A randomly selected subset of the Ames data set. The dataset has had its values imputed and any remaining NA values removed.

Usage

```
AmesImp
```

Format

A data frame with 2047 observations on 74 variables.

Source

https://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt

Examples

```
str(AmesImp)
plot(AmesImp$Neighborhood, y = AmesImp$SalePrice)
```

`AmesImpFctrs` *A Subset of the Ames Data Set with Imputed Values Only Including Factor Variables and Sale Price*

Description

A randomly selected subset of the Ames data set. The dataset contains only factor class variables and the SalePrice variable. Missing values have been imputed.

Usage

```
AmesImpFctrs
```

Format

A data frame with 2049 observations on 39 variables.

Source

https://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt

Examples

```
str(AmesImpFctrs)
plot(AmesImpFctrs$Neighborhood, y = AmesImpFctrs$SalePrice)
```

AmesSubset *A Subset of the Ames Data Set*

Description

A randomly selected subset of the Ames data set.

Usage

```
AmesSubset
```

Format

A data frame with 2049 observations on 82 variables.

Source

https://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt

Examples

```
str(AmesSubset)
plot(AmesSubset$Neighborhood, y = AmesSubset$SalePrice)
```

bin.oth *Recategorization of Variables by Mapping Tables Within a List*

Description

The functions purpose is to recategorize a data.frame's variables by the elements identified in a list. Each element of the list must contain two columns. The first column contains the original values, and the second column contains the new values. The first column name of each element of the list must be a variable name in the data.frame. Effectively, each element of the list is a mapping table. The list generated from the tree.bins() function can be directly passed as an element to this function.

Usage

```
bin.oth(list, data)
```

Arguments

list	A list generated from the tree.bins() function or created by the user to the specifications laid out in the description.
data	A data.frame.

See Also

tree.bins, fct_relevel, factor, left_join

Examples

```
#Allows the user to generate a list from the tree.bins() function
sample.df <- AmesImpFctrs[, c("Neighborhood", "MS.Zoning", "SalePrice")]
lookup.list <- tree.bins(data = sample.df, y = SalePrice, return = "lkup.list")

#Create a new data.frame and use the created list to map recategorize its values
new.df <- head(AmesImpFctrs[, c("Neighborhood", "MS.Zoning", "Lot.Shape", "SalePrice")], 100)
oth.binned.df <- bin.oth(list = lookup.list, data = new.df)
```

tree.bins *Recategorization of Factor Variables by Decision Tree Leaves*

Description

The function takes in a data set that contains categorical variable(s) and a response variable. It creates a decision tree by using one of the categorical variables (class factor) and the response variable. The decision tree is created from the rpart() function from the 'rpart' package The rules from the leaves of the decision tree are extracted, and used to recategorize the appropriate categorical variable (predictor). This step is performed for each of the categorical (class factor) variables that is fed into the data component of the function. Only variables containing more than 2 factors will be considered in the function. The final output generates a data set containing the recategorized variables or a list containing a mapping table for each of the candidate variables.

Usage

```
tree.bins(data, y, bin.nm = "Group.", method = NULL, control = NULL,
  return = "new.fctrs")
```

Arguments

data	A data.frame.
y	The response variables to be used in the rpart() function.
bin.nm	The string that will be used to categorize the variables. The default "Group." will be assigned. E.g. If a variable of 6 factors is recategorized into 3 factors, then setting bin.name equal to "Group." will name the three new factors to "Group.1", "Group.2", and "Group.3"
method	This is the method that will be used in the rpart() function. If null, the default method will be used. See rpart() for further detail.
control	This is the control that will be used in the rpart() function. The user has 3 options, one of which is the default selected control by the rpart() function. The remaining two option are: 1) Specity a cp value which will prune each decision tree by the specified value or 2) Specity a two-dimensional data.frame() that contains the variable name(s) as identified in the data component for the first column and the respective cp of each variable in the second column. Variable(s) not included in this data.frame() will use the cp generated by the rpart() function. See rpart() and rpart.control() for further detail.
return	This is what the function will return. There are three options: 1) new.fctrs - will provide a data.frame with the recategorized categorical variables. 2) lkup.list - will provide a list of lookup tables. Each element will contain the original to new mapping for each recategorized variable. 3) both - it will return both: the new.fctrs and lkup.list objects.

See Also

bin.oth, rpart, rpart.control,rpart.lists

Examples

```
#Returns a data.frame of recategorized variables
library(rpart)
sample.df <- AmesImpFctrs[, c("Neighborhood", "MS.Zoning", "SalePrice")]
tree.bins(data = sample.df, y = SalePrice)

#Returns a list of mapping tables generated from tree.bins()
tree.bins(data = sample.df, y = SalePrice, return = "lkup.list")

#Allows the user to choose the naming convention for the attribute naming convention
tree.bins(data = sample.df, y = SalePrice, bin.nm = "bin#")

#Allows user to manually assign a cp to each decision tree evaluated in rpart()
tree.bins(data = sample.df, y = SalePrice, control = rpart.control(cp = .01))
```

```
#Allows user to manually assign a cp to specified variables
demo.df <- data.frame(Variables = c("Neighborhood", "MS.Zoning"), CP = c(.001, .2))
tree.bins(data = sample.df, y = SalePrice, control = demo.df)
```

Index