Package 'smbinning'

June 14, 2015

Title Optimal Binning for Scoring Modeling

Version 0.2

Author Herman Jopia

Maintainer Herman Jopia <hjopia@gmail.com>

URL http://www.scoringmodeling.com/rpackage/smbinning

Description It categorizes a numeric variable into bins mapped to a binary target variable for its ulterior usage in scoring modeling. Its purpose is to automate the time consuming process of selecting the right cut points, quickly calculate metrics such as Weight of Evidence and Information Value; and also document SQL codes, tables, and plots used throughout the development stage. The package also allows users to establish their own cut points for numeric characteristics and run the analysis for categorical variables.

Depends R (>= 3.1.2),sqldf,partykit,Formula

Imports gsubfn

License GPL (>= 2)

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2015-06-14 01:25:19

R topics documented:

`chileancredit` *Chilean Credit Data*

Description

A simulated dataset based on six months of information collected by a Chilean Bank whose objective was to develop a credit scoring model to determine the probability of default within the next 12 months. The target variable is FlagGB, which represents the binary status of default (0) and not default(1).

Format

Data frame with 29,519 rows and 57 columns.

Details

- CustomerId. Customer Identifier.
- TOB. Time on books in months since first account was open.
- IncomeLevel. Income level from 00 (Low) to 05 (High).
- RevAccts. Number of open revolving accounts.
- InsAccs. Number of open installment accounts.
- RevBal. Outstanding balance in all open revolving accounts
- InsBal. Outstanding balance in all open installment accounts
- RevLim. Limit of all open revolving accounts
- SavingsAmtL3M. Amount saved in the last 3 months.
- Bal. Outstanding balance
- MaxDqBin. Max. delinquency bin. 0:No Dq., 1:1-29 ... 6:150-179.
- BureauBalDq1st. Outstanding balance 30-89 in Credit Bureau.
- BureauBalDq2nd.Outstanding balance 90-179 in Credit Bureau.
- BureauBalDq3rd.Outstanding balance 180+ in Credit Bureau.
- CntOtherLenders. Numer of other lenders (Credit Bureau).
- MtgBal. Mortgage outstanding balance at the Credit Bureau.
- NonBankTradesDq. Number of non-bank delinquent trades.
- Performance. Scoring model performance definition.
- CatGBI. Exclusion (NULL), Bad (00), Indet. (01), Good (02).
- FlagGB. 1: Good, 0: Bad.
- Random. Uniformly distributed random value for sampling purposes.
- FlagSample. Training and test sample indicator (1:75%,0:25%).

smbinning *Optimal Binning for Scoring Modeling*

Description

Optimal Binning categorizes a numeric characteristic into bins for ulterior usage in scoring modeling. This process, also known as *supervised discretization*, utilizes Recursive Partitioning to categorize the numeric characteristic.
The especific algorithm is Conditional Inference Trees which initially excludes missing values (NA) to compute the cutpoints, adding them back later in the process for the calculation of the *Information Value*.

Usage

```
smbinning(df, y, x, p = 0.05)
```

Arguments

df	A data frame.
y	Binary response variable (0,1). Integer (int) is required. Name of y must not have a dot.
x	Continuous characteristic. At least 10 different values. Value Inf is not allowed. Name of x must not have a dot.
p	Percentage of records per bin. Default 5% (0.05). This parameter only accepts values greater that 0.00 (0%) and lower than 0.5 (50%).

Value

The command smbinning generates and object containing the necessary info and utilities for binning. The user should save the output result so it can be used with smbinning.plot, smbinning.sql, and smbinning.gen.

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
str(chileancredit) # Quick description of the data
table(chileancredit$FlagGB) # Tabulate target variable

# Training and testing samples (Just some basic formality for Modeling)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)

# Package application
result=smbinning(df=chileancredit.train,y="FlagGB",x="TOB",p=0.05) # Run and save result
result$ivtable # Tabulation and Information Value
```

```
result$iv # Information value
result$bands # Bins or bands
result$ctree # Decision tree from partykit
```

`smbinning.custom` *Customized Binning*

Description

It gives the user the ability to create customized cutpoints. In Scoring Modeling, the analysis of a characteristic usually begins with intervals with the same length to understand its distribution, and then intervals with the same proportion of cases to explore bins with a reasonable sample size.

Usage

```
smbinning.custom(df, y, x, cuts)
```

Arguments

`df`	A data frame.
`y`	Binary response variable (0,1). Integer (`int`) is required. Name of `y` must not have a dot.
`x`	Continuous characteristic. At least 10 different values. Value `Inf` is not allowed. Name of `x` must not have a dot.
`cuts`	Vector with the cutpoints selected by the user. It does not have a default so user must define it.

Value

The command `smbinning.custom` generates and object containing the necessary info and utilities for binning. The user should save the output result so it can be used with `smbinning.plot`, `smbinning.sql`, and `smbinning.gen`.

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
str(chileancredit) # Quick description of the data
table(chileancredit$FlagGB) # Tabulate target variable

# Training and testing samples (Just some basic formality for Modeling)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)

# Remove exclusions from chileancredit dataset
TOB.train=
  subset(chileancredit,(FlagSample==1 & (FlagGB==1 | FlagGB==0)), select=TOB)
```

```
TOB.test=
  subset(chileancredit,(FlagSample==0 & (FlagGB==1 | FlagGB==0)), select=TOB)

# Custom cutpoints using percentiles (20% each)
TOB.Pct20=quantile(TOB.train, probs=seq(0,1,0.2), na.rm=TRUE)
TOB.Pct20.Breaks=as.vector(quantile(TOB.train, probs=seq(0,1,0.2), na.rm=TRUE))
Cuts.TOB.Pct20=TOB.Pct20.Breaks[2:(length(TOB.Pct20.Breaks)-1)]

# Package application and results
result=
  smbinning.custom(df=chileancredit.train,
                   y="FlagGB",x="TOB",cuts=Cuts.TOB.Pct20) # Run and save
result$ivtable # Tabulation and Information Value
```

`smbinning.factor` *Binning on Factor Variables*

Description

It generates the output table for the uniques values of a given factor variable.

Usage

```
smbinning.factor(df, y, x)
```

Arguments

`df`	A data frame.
`y`	Binary response variable (0,1). Integer (`int`) is required. Name of `y` must not have a dot.
`x`	A factor variable with at least 2 different values. Value `Inf` is not allowed. Name of `x` must not have a dot.

Value

The command `smbinning.factor` generates and object containing the necessary info and utilities for binning. The user should save the output result so it can be used with `smbinning.plot`, `smbinning.sql`, and `smbinning.gen`.

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
str(chileancredit) # Quick description of the data
table(chileancredit$FlagGB) # Tabulate target variable

# Data transformation. Data type must be factor.
chileancredit$IncomeLevel= factor(chileancredit$IncomeLevel,
```

```
                                levels=c(0,1,2,3,4,5),
                                labels=c("00","01","02","03","04","05"))

# Training and testing samples (Just some basic formality for Modeling)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)

# Package application and results
result.train=smbinning.factor(df=chileancredit.train,
                              y="FlagGB",x="IncomeLevel")
result.train$ivtable
result.test=smbinning.factor(df=chileancredit.test,
                              y="FlagGB",x="IncomeLevel")
result.test$ivtable

# Plots
par(mfrow=c(2,2))
smbinning.plot(result.train,option="dist",sub="Income Level (Tranining Sample)")
smbinning.plot(result.train,option="badrate",sub="Income Level (Tranining Sample)")
smbinning.plot(result.test,option="dist",sub="Income Level (Test Sample)")
smbinning.plot(result.test,option="badrate",sub="Income Level (Test Sample)")
```

`smbinning.gen` *Utility to generate a new characteristic*

Description

It generates a data frame with a new predictive characteristic after the binning process.

Usage

```
smbinning.gen(df, ivout, chrname = "NewChar")
```

Arguments

`df`	Dataset to be updated with the new characteristic.
`ivout`	An object generated after `smbinning`.
`chrname`	Name of the new characteristic.

Value

A data frame with the binned version of the characteristic analyzed with `smbinning`.

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)
result=smbinning(df=chileancredit.train,y="FlagGB",x="TOB",p=0.05) # Run and save result

# Generate new binned characteristic into a existing data frame
chileancredit.train=
smbinning.gen(chileancredit.train,result,"gTOB") # Update training sample
chileancredit=
  smbinning.gen(chileancredit,result,"gTOB") # Update population
sqldf("select gTOB,count(*) as Recs
      from chileancredit group by gTOB") # Check new field counts
```

---

smbinning.plot *Plots after binning*

Description

It generates plots for distribution, bad rate, and weight of evidence after running `smbinning` and saving its output.

Usage

```
smbinning.plot(ivout, option = "dist", sub = "")
```

Arguments

`ivout`	An object generated by binning.
`option`	Distribution ("dist"), Good Rate ("goodrate"), Bad Rate ("badrate"), and Weight of Evidence ("WoE").
`sub`	Subtitle for the chart (optional).

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)
result=smbinning(df=chileancredit.train,y="FlagGB",x="TOB",p=0.05) # Run and save result

# Plots
par(mfrow=c(2,2))
boxplot(chileancredit.train$TOB~chileancredit.train$FlagGB,
        horizontal=TRUE, frame=FALSE, col="lightgray",main="Distribution")
```

```
mtext("Time on Books (Months)",3)
smbinning.plot(result,option="dist",sub="Time on Books (Months)")
smbinning.plot(result,option="badrate",sub="Time on Books (Months)")
smbinning.plot(result,option="WoE",sub="Time on Books (Months)")
```

`smbinning.sql` *SQL Code*

Description

It outputs a SQL code to facilitate the generation of new binned characetristic in a SQL environment.

Usage

```
smbinning.sql(ivout)
```

Arguments

`ivout`	An object generated by `smbinning`.

Value

A text with the SQL code for binning.

Examples

```
# Package loading and data exploration
library(smbinning) # Load package and its data
data(chileancredit) # Load smbinning sample dataset (Chilean Credit)
chileancredit.train=subset(chileancredit,FlagSample==1)
chileancredit.test=subset(chileancredit,FlagSample==0)
result=smbinning(df=chileancredit.train,y="FlagGB",x="TOB",p=0.05) # Run and save result

# Generate SQL code
smbinning.sql(result)
```

Index