05.03.c-Feature-Analysis-HCluster-w-Encoded-HiCard-Vars.R

cmhenn

2020-03-06

library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(FactoMineR)  
library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(ClustOfVar)  
  
# set local working directory  
setwd("~/Documents/GitHub/Incident-Management-Process-BPIC2014/")  
  
### for ChurnDataAsIs  
# import data from source file  
SLAData <- read.csv("data/05.00 Incident Data Encoded HiCard Vars.csv")  
  
# Switch SLAFail from 0 and 1 to No and Yes  
SLAData$SLAFail <- as.factor(recode(SLAData$SLAFail, '0' = "No", '1' = "Yes"))  
  
# Set Open\_Time\_HourOfDay to factor  
SLAData$Open\_Time\_HourOfDay <- as.factor(as.character(SLAData$Open\_Time\_HourOfDay))  
  
# summarize the data  
str(SLAData)

## 'data.frame': 35208 obs. of 10 variables:  
## $ Service\_Component\_WBS\_aff : num -0.59 0.636 0.551 -0.781 0.736 ...  
## $ Urgency : Factor w/ 5 levels "1 Very High",..: 4 5 5 5 5 3 4 3 4 4 ...  
## $ KM\_number : num -1.944 0.637 1.259 0 -0.304 ...  
## $ Count\_Related\_Interactions: num 1 1 1 1 1 3 1 1 1 1 ...  
## $ Count\_Related\_Incidents : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Count\_Related\_Changes : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SLAFail : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 1 1 1 ...  
## $ Open\_Time\_HourOfDay : Factor w/ 24 levels "0","1","10","11",..: 22 23 23 23 23 23 24 24 24 23 ...  
## $ Open\_Time\_DayOfWeek : Factor w/ 7 levels "Friday","Monday",..: 6 6 6 6 6 6 6 6 6 6 ...  
## $ CI\_TypeSubType\_aff : num -0.851 0.627 -0.12 -0.12 0.842 ...

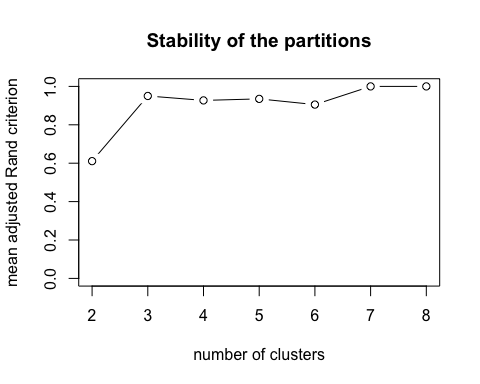
# check for null values  
apply(is.na(SLAData), 2, which)

## integer(0)

quantPredVars <- c('Service\_Component\_WBS\_aff',  
 'KM\_number',  
 'Count\_Related\_Interactions',  
 'Count\_Related\_Incidents',  
 'Count\_Related\_Changes',  
 'CI\_TypeSubType\_aff'  
 )  
# SLAData$Open\_Time\_HourOfDay < as.factor(SLAData$Open\_Time\_HourOfDay)  
qualPredVars <- c('Urgency',  
 'Open\_Time\_DayOfWeek',  
 'Open\_Time\_HourOfDay')  
  
  
### variable clustering  
variable\_tree <- hclustvar(X.quanti = subset(SLAData, select = quantPredVars),  
 X.quali = subset(SLAData, select = qualPredVars),  
 )  
png(filename = "reports/05.03.c hclust dendogram.png", height = 450, width = 800)  
plot(variable\_tree)  
dev.off()

## quartz\_off\_screen   
## 2

stability(variable\_tree, B = 25)



##   
## Call:  
## stability(tree = variable\_tree, B = 25)  
##   
##   
## name description   
## "$matCR" "matrix of corrected Rand indices"   
## "$meanR" "vector of mean corrected Rand indices"

clus3 <- cutreevar(variable\_tree, 3)  
clus3$var

## $cluster1  
## squared loading correlation  
## Service\_Component\_WBS\_aff 0.7197745 0.8483953  
## KM\_number 0.6929269 0.8324223  
## CI\_TypeSubType\_aff 0.5129899 0.7162331  
##   
## $cluster2  
## squared loading correlation  
## Count\_Related\_Incidents 0.57711449 0.7596805  
## Urgency 0.39516986 NA  
## Count\_Related\_Interactions 0.33524712 0.5790053  
## Count\_Related\_Changes 0.06451964 0.2540072  
##   
## $cluster3  
## squared loading correlation  
## Open\_Time\_DayOfWeek 0.7262746 NA  
## Open\_Time\_HourOfDay 0.7262746 NA