Capstone_Project_Report_Miguel_Sanchez

Miguel Sanchez

28-09-2020

INTRODUCTION

Machine Learning is a subset of Data Science and it´s becoming a strategic piece of digital transformation processes.

Predictive algorithms provide additional insights to make better decisions and will enable proactive actions on a particular business pain point.

The current initiative is intended for the final EDX-Capstone, implementing a Machine Learning platform for fraud detection.

This report is assembled with four sections:

The CONTEXT section provideS the business pain point, goal and objetives for the predictive platform and also will detail the data set used for training, test and validation purposes.

The METHOD /ANALYSIS section provides the data transformation and cleaning techniques as well as data balance methods.The METHOD will also cover the different Machine Learning algorithms used in the platform.

The RESULTS section provides the output on each tested algorithm and also provide details on the execution of the selected algorithm against the validation data set.

The CONCLUSION section provides recommendations, lessons learned and next steps related with the platform.

1) CONTEXT

Fraud and Risk are relevant topics for Banks and financial institutions; most of the current initiatives for fraud/risk mitigation have a reactive approach, triggering cutomer disappointment, frustation and having a direct impact on KPIS´s related with NPS (Customer Net Promote Score), CXI (Customer Experience Index) and overall customer satisfaction.

Machine learning models to detect and prevent risky (an eventually fraudulent) transactions in predictive way, can provide the Banks a proactive approach, having the opportunity to react in advance, taking the proper mitigation actions.

1.1) THE PROBLEM

The Bank XX has deployed a web application, offering an inter-bank money tranfer service. Several complaints are being received from customers, stating a fraud (identity thief) was committed as they didn´t execute a money transfer transaction.

1.2) THE APPROACH

To create and deploy a machine learning platform that is able to proactively detect suspicious transactions; the transaction should be flagged (moved to "stand by" state and not executed) so the call center can contact the customer and validate for the intended transaction.

1.3) THE DATA

I have implemented a system to detect risky transactions, using synthetic data for train and validate the model.

Base Data Set: 6.681.203 Records

Base Data Set: 117 variables

Data set has been created using synthetic methods. Real/Transactional data used as a seed, coming from a Banking legacy/core platform.

Most or the variables are categorical as data is coming from a tranasactional system, only a few of them are continuous (i.e. balance, deposit)

CLASS variable used for training and prediction –> 0 for regular txn´s –> 1 for suspicious txn´s that could lead on a FRAUD

Values for the CLASS variable were assigned based on real occurrences of suspicious vs not suspicious transactions, using a Data Engineering process.

Additional variables with the prediction will be created over the test data set, depending on the used algorithm.

```
nrow(base)
```

```
## [1] 6681203
```

```
str(base, list.len=ncol(base))
```

```
## 'data.frame':    6681203 obs. of  117 variables:
##  $ customer_id       : chr  "1" "1" "1" "1" ...
##  $ FROM              : chr  "20190206" "20190610" "20190626" "20190628" ...
##  $ UPTO              : chr  "20190211" "20190615" "20190701" "20190703" ...
##  $ ACANXCLADIN       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACTDATSMS         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ APVP2             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ APVP3             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ AVCETRAM          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BLOQCLACC         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BLOQCLACCE        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BUSQRUT           : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 2 2 2 ...
##  $ CAND1             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CAND2             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CAVP2             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CAVP2AUT          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CAVP3             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CCOTIZ            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CDENROLWEB        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ CDMODCEL          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CDREVDESENR       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CERTAFIL          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CHECK             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINCONF        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINDESENR      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINENR1        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINMOD1        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINMOD2        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINP1          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINP2          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINP3          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINREV1        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLADINREV2        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CLSEGREST         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSCLASEG        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSEXPELEC       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSHISTANT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ CONSPCLI          : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 2 2 2 ...
##  $ CONSREGAT         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSSALDOCCV      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSTRAM          : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ CONSVALCERT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CONSVINLAB        : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1 1 ...
##  $ CONTSEREMOT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CPP               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CRIM              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DASHBENEF         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DEPOSIT           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EFECTI            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EMAILACTDAT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EMAILCREACL       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EMAILMODDAT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EMAILRECCLACC     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ENTCLASEGAD       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ENVCLAEMAIL       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ GENCLAVDIN        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ICOM              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2 1 ...
##  $ IDOPER            : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ IPRODSAL          : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 2 ...
##  $ INGAPP            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ INSCTABANC        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LINKCLASEG        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MAILPAGPENS       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MANDATE           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MCLAACCCLI        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MCLAACCFOR        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MCLASATFOR        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODALCLADIN       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODALCLADIN2      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODANTCLI         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODCEL            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODDIRCOM         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODDIROTR         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ MODDIRPAR         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODEMAILCOM       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODEMAILOTR       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODEMAILPAR       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODFONCOM         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODFONINT         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MODFONPAR         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OPECLADIN         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RECACCWEB         : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 1 1 ...
##  $ RECUPCLACCE       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RECUPCLIVR        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REPAVTRAM         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RESCLASEG         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RESCLASEGD        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RESSALDO          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RETCAV            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REVCONTCLASE      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BALANCE           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SECLACCFALL       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SECLSEG           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SMSMODCEL1        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SMSMODCEL2        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SOLEXCLISTP       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REQVP             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REQVPA            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WVP               : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ WVPA              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SOLRETCCV         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTPEMAIL         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTPSMS           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALCLI1           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALCLI2           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALCLI3           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALCLI4           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALCLI5           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LOG               : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SIMDEFAULT        : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ target            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GENDER_ID         : Factor w/ 2 levels "1","2": NA NA NA NA NA NA 2 2 2 1 ...
##  $ CUSTOMER_AGE      : int  NA NA NA NA NA NA 40 40 40 52 ...
##  $ CUST_SERVICE_SCORE: int  NA NA NA NA NA NA 905 905 905 1310 ...
##  $ CUST_PROFIT_SCORE : int  NA NA NA NA NA NA 21689 21689 21689 -2905 ...
##  $ dominio           : Factor w/ 17063 levels "03.cl","0azar.cl",..: NA NA NA NA NA NA 7039 7039 7039
##  $ CUSTOMER_ID       : num  1e+00 1e+00 1e+00 1e+00 1e+07 ...
##  $ qxtotp            : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
head(base)
```

```
##   customer_id     FROM     UPTO ACANXCLADIN ACTDATSMS APVP2 APVP3 AVCETRAM
## 1           1 20190206 20190211           0         0     0     0        0
## 2           1 20190610 20190615           0         0     0     0        0
## 3           1 20190626 20190701           0         0     0     0        0
## 4           1 20190628 20190703           0         0     0     0        0
## 5    10000007 20190919 20190924           0         0     0     0        0
```

```
## 6    10000028 20190523 20190528           0         0     0     0        0
##   BLOQCLACC BLOQCLACCE BUSQRUT CAND1 CAND2 CAVP2 CAVP2AUT CAVP3 CCOTIZ
## 1         0          0       1     0     0     0        0     0      0
## 2         0          0       0     0     0     0        0     0      0
## 3         0          0       0     0     0     0        0     0      0
## 4         0          0       0     0     0     0        0     0      0
## 5         0          0       1     0     0     0        0     0      0
## 6         0          0       1     0     0     0        0     0      0
##   CDENROLWEB CDMODCEL CDREVDESENR CERTAFIL CHECK CLADINCONF CLADINDESENR
## 1          0        0           0        0     0          0            0
## 2          0        0           0        0     0          0            0
## 3          0        0           0        0     0          0            0
## 4          0        0           0        0     0          0            0
## 5          0        0           0        0     0          0            0
## 6          0        0           0        0     0          0            0
##   CLADINENR1 CLADINMOD1 CLADINMOD2 CLADINP1 CLADINP2 CLADINP3 CLADINREV1
## 1          0          0          0        0        0        0          0
## 2          0          0          0        0        0        0          0
## 3          0          0          0        0        0        0          0
## 4          0          0          0        0        0        0          0
## 5          0          0          0        0        0        0          0
## 6          0          0          0        0        0        0          0
##   CLADINREV2 CLSEGREST CONSCLASEG CONSEXPELEC CONSHISTANT CONSPCLI CONSREGAT
## 1          0         0          0           0           0        0         0
## 2          0         0          0           0           0        0         0
## 3          0         0          0           0           0        0         0
## 4          0         0          0           0           0        0         0
## 5          0         0          0           0           0        1         0
## 6          0         0          0           0           1        1         0
##   CONSSALDOCCV CONSTRAM CONSVALCERT CONSVINLAB CONTSEREMOT CPP CRIM DASHBENEF
## 1            0        0           0          0           0   0    0         0
## 2            0        0           0          0           0   0    0         0
## 3            0        0           0          0           0   0    0         0
## 4            0        0           0          0           0   0    0         0
## 5            0        0           0          1           0   0    0         0
## 6            0        1           0          1           0   0    0         0
##   DEPOSIT EFECTI EMAILACTDAT EMAILCREACL EMAILMODDAT EMAILRECCLACC ENTCLASEGAD
## 1       0      0           0           0           0             0           0
## 2       0      0           0           0           0             0           0
## 3       0      0           0           0           0             0           0
## 4       0      0           0           0           0             0           0
## 5       0      0           0           0           0             0           0
## 6       0      0           0           0           0             0           0
##   ENVCLAEMAIL GENCLAVDIN ICOM IDOPER IPRODSAL INGAPP INSCTABANC LINKCLASEG
## 1           0          0    0      0        0      0          0          0
## 2           0          0    0      0        0      0          0          0
## 3           0          0    0      0        0      0          0          0
## 4           0          0    0      0        0      0          0          0
## 5           0          0    0      0        0      0          0          0
## 6           0          0    0      1        1      0          0          0
##   MAILPAGPENS MANDATE MCLAACCCLI MCLAACCFOR MCLASATFOR MODALCLADIN MODALCLADIN2
## 1           0       0          0          0          0           0            0
## 2           0       0          0          0          0           0            0
## 3           0       0          0          0          0           0            0
```

```
## 4           0       0          0          0          0           0            0
## 5           0       0          0          0          0           0            0
## 6           0       0          0          0          0           0            0
##   MODANTCLI MODCEL MODDIRCOM MODDIROTR MODDIRPAR MODEMAILCOM MODEMAILOTR
## 1         0      0         0         0         0           0           0
## 2         0      0         0         0         0           0           0
## 3         0      0         0         0         0           0           0
## 4         0      0         0         0         0           0           0
## 5         0      0         0         0         0           0           0
## 6         0      0         0         0         0           0           0
##   MODEMAILPAR MODFONCOM MODFONINT MODFONPAR OPECLADIN RECACCWEB RECUPCLACCE
## 1           0         0         0         0         0         0           0
## 2           0         0         0         0         0         1           0
## 3           0         0         0         0         0         1           0
## 4           0         0         0         0         0         1           0
## 5           0         0         0         0         0         0           0
## 6           0         0         0         0         0         0           0
##   RECUPCLIVR REPAVTRAM RESCLASEG RESCLASEGD RESSALDO RETCAV REVCONTCLASE
## 1          0         0         0          0        0      0            0
## 2          0         0         0          0        0      0            0
## 3          0         0         0          0        0      0            0
## 4          0         0         0          0        0      0            0
## 5          0         0         0          0        0      0            0
## 6          0         0         0          0        0      0            0
##   BALANCE SECLACCFALL SECLSEG SMSMODCEL1 SMSMODCEL2 SOLEXCLISTP REQVP REQVPA
## 1       0           0       0          0          0           0     0      0
## 2       0           0       0          0          0           0     0      0
## 3       0           0       0          0          0           0     0      0
## 4       0           0       0          0          0           0     0      0
## 5       0           0       0          0          0           0     0      0
## 6       0           0       0          0          0           0     0      0
##   WVP WVPA SOLRETCCV TOTPEMAIL TOTPSMS VALCLI1 VALCLI2 VALCLI3 VALCLI4 VALCLI5
## 1   0    0         0         0       0       0       0       0       0       0
## 2   0    0         0         0       0       0       0       0       0       0
## 3   0    0         0         0       0       0       0       0       0       0
## 4   0    0         0         0       0       0       0       0       0       0
## 5   0    0         0         0       0       0       0       0       0       0
## 6   0    0         0         0       0       0       0       0       0       0
##   LOG SIMDEFAULT target GENDER_ID CUSTOMER_AGE CUST_SERVICE_SCORE
## 1   0          0      0      <NA>           NA                 NA
## 2   0          0      0      <NA>           NA                 NA
## 3   0          0      0      <NA>           NA                 NA
## 4   0          0      0      <NA>           NA                 NA
## 5   0          0      0      <NA>           NA                 NA
## 6   0          0      0      <NA>           NA                 NA
##   CUST_PROFIT_SCORE dominio CUSTOMER_ID qxtotp
## 1                NA    <NA>           1      0
## 2                NA    <NA>           1      0
## 3                NA    <NA>           1      0
## 4                NA    <NA>           1      0
## 5                NA    <NA>    10000007      0
## 6                NA    <NA>    10000028      0
```

1.3.1) FEATURE SELECTION & DATA WRANGLING

Initial analysis will be performed over the original data set to determine relevant variables (not all the variables should be used). Machine Learning algorithms perform better if highly correlated attributes are removed.

DATA WRANGLING - deleting LOG´s and SIMDEFAULT variables as those are systemic.Setting TARGET variable to CLASS and deleting TARGET

```
base$SIMDEFAULT <- NULL
base$LOG <- NULL

base$class <- base$target
base$target <- NULL
```

1.3.1.1) VARIABLE REDUNDANCY - METHOD 1

The Caret R package provides the findCorrelation which will analyze a correlation matrix of my data's attributes report on attributes that can be removed. I want to remove attributes with an absolute correlation of (ideally >0.75).

```
set.seed(7)
```

Using a sample dataset to determine variable redundancy;only 1.000.000 records will be processed because of memory limitations

```
basep <- base[sample(1:6340852,1000000),4:107]
basep[] <- lapply (basep, function (x) as.numeric (as.character (x)))
correlationMatrix <- cor(basep)
```

Summarize the correlation matrix

```
options(max.print=1000)
print(correlationMatrix)
```

```
##                 ACANXCLADIN     ACTDATSMS         APVP2         APVP3
## ACANXCLADIN    1.000000e+00  1.170283e-02  0.0631404332  0.0591650218
## ACTDATSMS      1.170283e-02  1.000000e+00  0.0075892898  0.0070143256
## APVP2          6.314043e-02  7.589290e-03  1.0000000000  0.9701832831
## APVP3          5.916502e-02  7.014326e-03  0.9701832831  1.0000000000
## AVCETRAM       4.353129e-02  5.199287e-03  0.0836290594  0.0811156700
## BLOQCLACC      3.503838e-02  8.469938e-03  0.0129546177  0.0112272812
## BLOQCLACCE     7.603813e-03  3.157152e-03 -0.0006994488 -0.0004488484
## BUSQRUT        4.075728e-02  4.922278e-02  0.0054463920  0.0060078559
## CAND1          4.786748e-02  7.390760e-02  0.0180194598  0.0176577530
##                     AVCETRAM     BLOQCLACC    BLOQCLACCE       BUSQRUT
## ACANXCLADIN    0.0435312858  3.503838e-02  7.603813e-03  4.075728e-02
## ACTDATSMS      0.0051992868  8.469938e-03  3.157152e-03  4.922278e-02
## APVP2          0.0836290594  1.295462e-02 -6.994488e-04  5.446392e-03
## APVP3          0.0811156700  1.122728e-02 -4.488484e-04  6.007856e-03
## AVCETRAM       1.0000000000  1.799866e-02  4.622902e-03  3.547294e-02
## BLOQCLACC      0.0179986572  1.000000e+00  3.001721e-01 -3.735076e-02
```

```
## BLOQCLACCE     0.0046229016  3.001721e-01  1.000000e+00 -1.463924e-02
## BUSQRUT        0.0354729355 -3.735076e-02 -1.463924e-02  1.000000e+00
## CAND1          0.0238818525  4.841216e-02  7.405219e-03  1.819042e-01
##                       CAND1         CAND2         CAVP2      CAVP2AUT
## ACANXCLADIN    0.047867482  0.1762791574  0.0506987352  3.062839e-02
## ACTDATSMS      0.073907602  0.0531398492  0.0021066431  5.559772e-03
## APVP2          0.018019460  0.0538035584  0.1142320559  8.146708e-03
## APVP3          0.017657753  0.0523713819  0.1118507286  7.184356e-03
## AVCETRAM       0.023881852  0.0426354767  0.0854809199  4.503799e-02
## BLOQCLACC      0.048412156  0.0297435201  0.0071418759  9.049927e-03
## BLOQCLACCE     0.007405219  0.0059342440 -0.0009674264 -1.289154e-04
## BUSQRUT        0.181904161  0.2295613960  0.0006411588  2.209289e-02
## CAND1          1.000000000  0.2751511135  0.0162958979  2.405036e-02
##                       CAVP3        CCOTIZ    CDENROLWEB      CDMODCEL
## ACANXCLADIN    0.0460197779  7.122478e-03  9.493633e-01  1.745561e-02
## ACTDATSMS      0.0019506095  7.082602e-03  1.193572e-02  1.406617e-02
## APVP2          0.1126659208 -1.733700e-02  6.270793e-02  1.988112e-02
## APVP3          0.1128999585 -1.687994e-02  5.854468e-02  1.956736e-02
## AVCETRAM       0.0828490765  1.204076e-04  4.338159e-02  8.053342e-03
## BLOQCLACC      0.0062646744  4.309224e-02  3.405489e-02  5.686479e-03
## BLOQCLACCE    -0.0009530849 -6.964288e-03  7.193474e-03  3.453517e-04
## BUSQRUT        0.0016073834 -2.052845e-02  3.923038e-02  7.391576e-03
## CAND1          0.0160817247  4.839335e-02  4.894733e-02  9.047899e-03
##                  CDREVDESENR      CERTAFIL         CHECK    CLADINCONF
## ACANXCLADIN    6.185584e-02  2.177426e-03 -1.597986e-04  1.091050e-02
## ACTDATSMS      1.611582e-02  1.672118e-02 -5.341085e-05  9.209122e-03
## APVP2          2.668432e-02 -2.780649e-03 -2.263754e-04  1.669706e-02
## APVP3          2.434299e-02 -2.647568e-03 -2.196256e-04  1.611015e-02
## AVCETRAM       1.745095e-02 -1.521597e-03 -2.977000e-04  4.341768e-03
## BLOQCLACC      9.000119e-03  2.278941e-03 -2.887084e-04  4.936612e-04
## BLOQCLACCE     2.342360e-03 -1.564293e-04 -9.353860e-05 -7.286262e-04
## BUSQRUT        1.233546e-02  4.419163e-02  2.781023e-03  3.326072e-03
## CAND1          1.382678e-02  1.063914e-01  4.940922e-03  6.381018e-03
##                CLADINDESENR    CLADINENR1    CLADINMOD1    CLADINMOD2
## ACANXCLADIN    1.034975e-01  5.895753e-01  2.127437e-05  1.265337e-05
## ACTDATSMS      1.415939e-02  2.508298e-03  5.695825e-03  5.668583e-03
## APVP2          2.353909e-02  3.289236e-02  1.593411e-02  1.585379e-02
## APVP3          2.074572e-02  3.015902e-02  1.574737e-02  1.566821e-02
## AVCETRAM       1.199548e-02  2.353575e-02  4.266804e-03  4.232642e-03
## BLOQCLACC      4.562264e-03  2.194438e-02 -7.776592e-05 -9.284390e-05
## BLOQCLACCE     2.750746e-03  3.219966e-03 -5.816795e-04 -5.841817e-04
## BUSQRUT        1.200826e-02  1.278679e-02  1.154566e-03  8.860092e-04
## CAND1          7.979367e-03  1.922174e-02  2.905691e-03  2.882725e-03
##                    CLADINP1      CLADINP2      CLADINP3    CLADINREV1
## ACANXCLADIN    8.846965e-01  9.584676e-01  6.406232e-01  4.660246e-02
## ACTDATSMS      8.365186e-03  1.081876e-02  8.960996e-03  9.095504e-03
## APVP2          6.387904e-02  6.513333e-02  5.403930e-02  1.068556e-02
## APVP3          6.025826e-02  6.127294e-02  5.178345e-02  7.210057e-03
## AVCETRAM       4.224986e-02  4.505997e-02  3.427625e-02  1.212053e-02
## BLOQCLACC      2.985727e-02  3.540486e-02  1.936346e-02  1.852431e-03
## BLOQCLACCE     3.896152e-03  7.251657e-03  1.559706e-03  4.827610e-03
## BUSQRUT        4.266183e-02  4.364012e-02  4.368690e-02  7.130110e-03
## CAND1          4.913635e-02  5.242396e-02  4.682520e-02  5.705597e-03
##                  CLADINREV2     CLSEGREST    CONSCLASEG   CONSEXPELEC
```

```
##                 CONSHISTANT      CONSPCLI     CONSREGAT  CONSSALDOCCV
## ACANXCLADIN    1.106529e-02  0.0436323635  0.0206655227  0.0226138808
## ACTDATSMS      8.164484e-03  0.0474932270  0.0161319883  0.0099098787
## APVP2          2.038428e-02  0.0682439138  0.0103863816  0.0778452640
## APVP3          2.037567e-02  0.0703412290  0.0091094911  0.0761379267
## AVCETRAM       9.677187e-03  0.0500415823  0.0199908838  0.0208459928
## BLOQCLACC     -1.563377e-03 -0.0362037794 -0.0010216179 -0.0019496723
## BLOQCLACCE    -2.564226e-04 -0.0152629302 -0.0018355105 -0.0020829940
## BUSQRUT        4.564305e-02  0.9589949361  0.1581514191  0.1689594495
## CAND1          2.340443e-02  0.1753048730  0.0473187628  0.0467745283
##                    CONSTRAM   CONSVALCERT    CONSVINLAB   CONTSEREMOT
## ACANXCLADIN    0.0206988085  8.227771e-03  8.552239e-03  0.2341613027
## ACTDATSMS      0.0302009518  4.156476e-02  1.386606e-02  0.0301553837
## APVP2          0.0220717227 -5.101176e-03  3.893502e-03  0.0593866625
## APVP3          0.0216913456 -4.815602e-03  4.331172e-03  0.0575055775
## AVCETRAM       0.0670559803  1.226519e-02  2.515308e-03  0.0563250796
## BLOQCLACC     -0.0195593806 -8.268526e-03 -1.915153e-02  0.0504989467
## BLOQCLACCE    -0.0075868874 -5.310117e-03 -6.144696e-03  0.0209799827
## BUSQRUT        0.4853442665  3.617488e-01  3.251826e-01  0.1073339889
## CAND1          0.0723786181  1.691282e-01  3.557633e-02  0.1217137345
##                         CPP          CRIM     DASHBENEF       DEPOSIT
## ACANXCLADIN   -7.026488e-04  5.610774e-03  0.0406658962  1.305076e-02
## ACTDATSMS      4.027440e-03  7.417050e-03  0.0146875339  1.487601e-03
## APVP2         -9.953929e-04 -7.422862e-03  0.0091116925  1.366973e-01
## APVP3         -9.657136e-04 -7.624999e-03  0.0074272381  1.396887e-01
## AVCETRAM       1.050629e-03  2.029233e-02  0.0754245045  1.841525e-02
## BLOQCLACC      1.969162e-03  1.269901e-02  0.0487066197 -1.421516e-04
## BLOQCLACCE     2.027264e-03 -1.753593e-03  0.0066700380  1.268652e-03
## BUSQRUT        1.047271e-02  2.306455e-02 -0.0114270645 -1.454639e-06
## CAND1          2.133615e-02  2.569648e-02  0.0187202801  5.125202e-03
##                       EFECTI   EMAILACTDAT   EMAILCREACL   EMAILMODDAT
## ACANXCLADIN    5.496191e-04  3.704983e-02  0.0983789328  0.0662776521
## ACTDATSMS     -2.561579e-04  2.724001e-01  0.0167689954  0.0249874027
## APVP2          2.514661e-02  8.066133e-03  0.0245726082  0.0312498310
## APVP3          2.595850e-02  8.113798e-03  0.0226054570  0.0299064254
## AVCETRAM       5.062453e-03  1.235750e-02  0.0410182083  0.0537929011
## BLOQCLACC     -1.384642e-03  2.610296e-02  0.3751537668  0.0235325428
## BLOQCLACCE    -4.486102e-04  4.490590e-03  0.1122313990  0.0043935946
## BUSQRUT       -6.128604e-04  1.523585e-01 -0.0517541156  0.0112122287
## CAND1          2.268672e-03  2.836639e-01  0.1666077385  0.0419301271
##                EMAILRECCLACC   ENTCLASEGAD   ENVCLAEMAIL    GENCLAVDIN
## ACANXCLADIN    0.0108032357  0.3137811378  1.463804e-02  0.3145959755
## ACTDATSMS      0.0033318494  0.0362803008  7.064268e-02  0.0026003657
## APVP2         -0.0008904638  0.0864931942  3.416075e-03  0.2962976136
## APVP3         -0.0010462670  0.0840621311  3.573170e-03  0.2861337303
```

```
## AVCETRAM       0.0041042861  0.0607707424  6.822527e-03  0.1069844943
## BLOQCLACC      0.1816950857  0.0418291220  2.726965e-02  0.0342782925
## BLOQCLACCE     0.5813828859  0.0113819088  3.405598e-03  0.0088751794
## BUSQRUT       -0.0198637666  0.1363151989  7.869071e-02 -0.0474775109
## CAND1          0.0099582808  0.1796524481  4.309117e-01  0.0118288857
##                        ICOM        IDOPER     IPRODSAL        INGAPP
## ACANXCLADIN    0.0782879467  0.0025081800  0.052675471  1.132312e-02
## ACTDATSMS      0.0452508024  0.0255873669  0.049585874 -3.658027e-03
## APVP2          0.0316404335 -0.0066301589  0.023287086  1.065163e-02
## APVP3          0.0309580943 -0.0064935838  0.023281212  9.888982e-03
## AVCETRAM       0.0642661807  0.0196803005  0.045526123 -4.406291e-03
## BLOQCLACC      0.0083988649 -0.0192161056 -0.019917367 -1.358534e-02
## BLOQCLACCE    -0.0028637587 -0.0071153094 -0.008220742  6.810229e-02
## BUSQRUT        0.4907258509  0.3469823775  0.694411996 -1.661590e-01
## CAND1          0.2727880219  0.0562076466  0.162079654 -2.668109e-02
##                   INSCTABANC    LINKCLASEG   MAILPAGPENS       MANDATE
## ACANXCLADIN    5.583799e-02  0.0529785006 -8.511895e-05  1.316517e-01
## ACTDATSMS      2.267230e-03  0.0282271977  1.716680e-03  1.121046e-02
## APVP2          7.721706e-02  0.0188985089 -2.776582e-03  2.068847e-01
## APVP3          7.526476e-02  0.0181590607 -2.817328e-03  2.015836e-01
## AVCETRAM       4.859423e-02  0.0205566138  2.131794e-02  7.647144e-02
## BLOQCLACC      6.904339e-03  0.0317262838  3.908694e-03  1.469963e-02
## BLOQCLACCE    -1.642352e-03  0.0141273501  2.455395e-04  2.832750e-03
## BUSQRUT        2.022510e-02  0.0703521118  4.721446e-03  4.235158e-02
## CAND1          1.176946e-02  0.0998261064  1.261856e-02  4.373201e-02
##                   MCLAACCCLI    MCLAACCFOR    MCLASATFOR   MODALCLADIN
## ACANXCLADIN    2.691217e-02 -1.304750e-04 -1.304750e-04  1.107808e-01
## ACTDATSMS      5.851739e-03 -4.360975e-05 -4.360975e-05 -2.274013e-03
## APVP2          8.773926e-03 -1.848347e-04 -1.848347e-04  2.532784e-02
## APVP3          8.244118e-03 -1.793235e-04 -1.793235e-04  2.490664e-02
## AVCETRAM       1.911654e-02 -2.430709e-04 -2.430709e-04  6.545184e-02
## BLOQCLACC      3.684081e-02 -2.357293e-04 -2.357293e-04  3.349808e-02
## BLOQCLACCE     4.999337e-03 -7.637391e-05 -7.637391e-05 -2.885927e-03
## BUSQRUT        1.132503e-02  2.270695e-03  2.270695e-03 -1.642604e-01
## CAND1          2.646667e-02  1.242467e-02  1.242467e-02  1.276986e-02
##                  MODALCLADIN2     MODANTCLI        MODCEL     MODDIRCOM
## ACANXCLADIN    0.0629192068  0.0667381799  0.0287983321  3.551179e-03
## ACTDATSMS     -0.0023464342  0.1826326970  0.1360227312  2.100195e-02
## APVP2          0.0240434215  0.0213977917  0.0085491338  1.641271e-03
## APVP3          0.0237054882  0.0207938444  0.0084548050  1.208559e-03
## AVCETRAM       0.0657656616  0.0408859594  0.0092400243  1.169682e-03
## BLOQCLACC      0.0299379758  0.0182028794  0.0100404719  8.915881e-04
## BLOQCLACCE    -0.0049209058  0.0025400256  0.0028400962 -7.365911e-04
## BUSQRUT       -0.1614506759  0.1659862408  0.1322459905  1.977545e-02
## CAND1          0.0101807520  0.3027480636  0.2899983965  2.389873e-02
##                   MODDIROTR     MODDIRPAR   MODEMAILCOM   MODEMAILOTR
## ACANXCLADIN    5.600481e-03  0.0344110899  5.856132e-03  7.773056e-04
## ACTDATSMS      2.224922e-02  0.1780478943  3.275600e-02  7.927759e-03
## APVP2         -1.469499e-04  0.0072423086  8.103102e-04  3.101964e-03
## APVP3         -7.594209e-04  0.0074840295  9.709632e-04  3.285552e-03
## AVCETRAM       2.196326e-03  0.0150195793  3.810392e-03  3.590075e-03
## BLOQCLACC      7.683880e-04  0.0058462364  2.367825e-04 -1.433326e-03
## BLOQCLACCE    -6.205049e-04 -0.0001751547 -9.479203e-04 -6.344525e-04
## BUSQRUT        1.689663e-02  0.1415875103  2.551597e-02  1.715570e-02
```

```
## CAND1          2.657934e-02  0.2375771644  4.719736e-02  3.654370e-02
##                  MODEMAILPAR     MODFONCOM     MODFONINT     MODFONPAR
## ACANXCLADIN    1.396524e-02  6.809089e-03 -3.195987e-04  1.836862e-02
## ACTDATSMS      1.156135e-01  3.612901e-02  1.863391e-02  7.814566e-02
## APVP2         -1.119107e-03  7.969780e-05 -4.527529e-04  1.171660e-03
## APVP3         -1.046551e-03 -1.814585e-04 -4.392532e-04  8.042978e-04
## AVCETRAM       2.553365e-03  4.470562e-03 -5.954026e-04  6.799990e-03
## BLOQCLACC      9.287984e-03 -1.535570e-03 -5.774193e-04  1.363877e-03
## BLOQCLACCE    -7.499864e-05  6.624797e-05 -1.870780e-04 -1.041039e-03
## BUSQRUT        1.235782e-01  4.489041e-02  4.275477e-03  8.991423e-02
## CAND1          3.213375e-01  5.122686e-02  4.743795e-03  1.204620e-01
##                    OPECLADIN     RECACCWEB   RECUPCLACCE    RECUPCLIVR
## ACANXCLADIN    1.938804e-01  0.0848758234  0.0773340295  0.0610574354
## ACTDATSMS     -1.167780e-03  0.0107896798  0.0130199608  0.0257103192
## APVP2          3.131612e-01  0.0160443912  0.0167733898  0.0145243179
## APVP3          3.017718e-01  0.0141641565  0.0147816940  0.0137959550
## AVCETRAM       9.324369e-02  0.0312631551  0.0313826756  0.0221343359
## BLOQCLACC      2.769533e-02  0.3883945037  0.3829271880  0.0598252704
## BLOQCLACCE     5.892854e-03  0.1217196366  0.0305745340  0.0166034603
## BUSQRUT       -4.345270e-02 -0.0726754671 -0.0666052366  0.0518198846
## CAND1          2.291760e-03  0.0793631498  0.1012308066  0.1509835660
##                   REPAVTRAM     RESCLASEG    RESCLASEGD      RESSALDO
## ACANXCLADIN    0.0247990989  0.1773817244  1.910794e-02  0.0317181632
## ACTDATSMS      0.0044325043  0.0190278420 -8.978296e-04 -0.0059946093
## APVP2          0.0660984860  0.0605807186  2.187504e-02  0.0191818411
## APVP3          0.0649654752  0.0600522738  2.164981e-02  0.0171934496
## AVCETRAM       0.6535775393  0.0337668893  1.126226e-02  0.0781803647
## BLOQCLACC      0.0084956897  0.0111648368 -4.853151e-03 -0.0051363656
## BLOQCLACCE     0.0029683346 -0.0035944429 -1.572372e-03 -0.0163920489
## BUSQRUT        0.0359220821  0.1132509979 -1.510824e-03 -0.1875363452
## CAND1          0.0168017161  0.1236891853  1.014628e-02 -0.0162780877
##                       RETCAV  REVCONTCLASE       BALANCE   SECLACCFALL
## ACANXCLADIN    0.0431299096  7.369908e-02  7.590657e-02  0.0457307883
## ACTDATSMS      0.0015375672  3.452167e-02 -5.594799e-03  0.0154406961
## APVP2          0.1071738504  2.694996e-02  1.022939e-01  0.0109294868
## APVP3          0.1074469313  2.478874e-02  9.962269e-02  0.0099847240
## AVCETRAM       0.0785620774  2.677760e-02  1.391151e-01  0.0161872209
## BLOQCLACC      0.0043630887  4.699857e-02  8.081776e-02  0.1499541999
## BLOQCLACCE    -0.0015130864  2.366234e-02  8.321373e-03  0.0889055624
## BUSQRUT       -0.0027868334  9.713988e-02 -3.560408e-01 -0.0025573371
## CAND1          0.0147806337  1.250042e-01  2.435836e-02  0.1144373337
##                     SECLSEG    SMSMODCEL1    SMSMODCEL2   SOLEXCLISTP
## ACANXCLADIN    0.0655181532  0.0624387415  0.0508635556  5.953668e-02
## ACTDATSMS      0.0109294860  0.1445072546  0.0551741086  1.320451e-02
## APVP2          0.2212719906  0.0235237233  0.0137093098  2.326813e-02
## APVP3          0.2165748209  0.0223417354  0.0131455755  2.246057e-02
## AVCETRAM       0.1296172048  0.0371800975  0.0206677278  3.995070e-02
## BLOQCLACC      0.0396910968  0.0399023046  0.0336320237  1.812365e-02
## BLOQCLACCE     0.0100746089  0.0083784334  0.0065759897  2.896402e-03
## BUSQRUT       -0.0042640749  0.0778740840  0.1354026360  1.187298e-02
## CAND1          0.0499825794  0.1832696414  0.2600965959  3.959261e-02
##                       REQVP        REQVPA           WVP          WVPA
## ACANXCLADIN    0.0345834098  5.653983e-02  5.607477e-02  1.213291e-03
## ACTDATSMS      0.0101027990  3.104338e-03  6.392829e-03  5.362306e-03
```

```
## APVP2          0.5643392962  1.253051e-01  9.276388e-01 -5.561796e-03
## APVP3          0.5511974866  1.204198e-01  9.556191e-01 -5.456468e-03
## AVCETRAM       0.0667356440  9.495228e-02  7.726202e-02  8.215229e-03
## BLOQCLACC      0.0123795458  8.576969e-03  9.057830e-03 -1.456270e-03
## BLOQCLACCE     0.0018310892  3.421635e-05 -1.014732e-03 -1.593024e-03
## BUSQRUT        0.0166295383 -5.294310e-03  1.420126e-03  9.591009e-02
## CAND1          0.0283973793  1.651476e-02  1.590675e-02  1.885342e-02
##                   SOLRETCCV     TOTPEMAIL       TOTPSMS       VALCLI1
## ACANXCLADIN    4.306615e-03  0.0944057622  0.1177786867  0.1111511961
## ACTDATSMS      4.560640e-03  0.0289575474  0.0271252859  0.0336033359
## APVP2          1.773017e-02  0.0402993179  0.0435940696  0.0525559426
## APVP3          1.576612e-02  0.0377957194  0.0415171651  0.0515678438
## AVCETRAM       8.868723e-03  0.0520776507  0.0511187467  0.0970317768
## BLOQCLACC     -2.313172e-03  0.0722748849  0.0648211688  0.0279540824
## BLOQCLACCE    -7.198933e-04  0.0214236602  0.0254634271  0.0075312719
## BUSQRUT        7.623679e-02  0.0305901909  0.0437686365  0.2236546305
## CAND1          1.859189e-02  0.0563375659  0.0965387450  0.1770332434
##                    VALCLI2       VALCLI3       VALCLI4       VALCLI5
## ACANXCLADIN    0.1137520184  5.112800e-02  5.155132e-03  0.0255935513
## ACTDATSMS      0.0293697109  1.741043e-02  3.488469e-02  0.0552304726
## APVP2          0.0525534896  2.747443e-02  2.024420e-05  0.0039174376
## APVP3          0.0515413442  2.689946e-02 -1.153673e-04  0.0035255966
## AVCETRAM       0.0989585519  4.793461e-02  9.265293e-03  0.0213707290
## BLOQCLACC      0.0211775890  2.073866e-02  2.282807e-02  0.0214162195
## BLOQCLACCE     0.0057285365  4.427506e-03  1.295792e-02  0.0034356092
## BUSQRUT        0.1961799331  9.969805e-02  1.002463e-01  0.2694541251
## CAND1          0.0794535658  8.865137e-02  1.383978e-01  0.4160801456
##  [ reached getOption("max.print") -- omitted 95 rows ]
```

Find attributes that are highly correlated

```
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.75)
```

Print indexes of highly correlated attributes

```
print(highlyCorrelated)
```

```
##  [1]  25  26  34  91  94   3  11   1  20  23  10  62  64  52  47  78  13   4 100
```

Variables that will be left out due to its high correlation

```
summary(basep[,highlyCorrelated])
```

```
##     CLADINP1           CLADINP2          CONSPCLI        SMSMODCEL2
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :0.0000   Median :0.00000
##  Mean   :0.007766   Mean   :0.00899   Mean   :0.2966   Mean   :0.01015
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      REQVPA            APVP2            CAVP2          ACANXCLADIN
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
```

```
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.03015   Mean   :0.0168   Mean   :0.02078   Mean   :0.00844
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##    CLADINCONF         CLADINMOD1           CAND2           MCLAACCFOR
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.00000   Min.   :0e+00
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0e+00
##  Median :0.000000   Median :0.000000   Median :0.00000   Median :0e+00
##  Mean   :0.000182   Mean   :0.000116   Mean   :0.02028   Mean   :2e-06
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0e+00
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1e+00
##   MODALCLADIN       GENCLAVDIN       EMAILCREACL      RECACCWEB
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.000   Median :0.0000
##  Mean   :0.2366   Mean   :0.07746   Mean   :0.115   Mean   :0.1237
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.000   Max.   :1.0000
##      CAVP3            APVP3            VALCLI1
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.01964   Mean   :0.01582   Mean   :0.02002
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
```

```
rm(basep)
```

1.3.1.2) VARIABLE REDUNDANCY - METHOD 2

Building a Learning Vector Quantization (LVQ) model. The varImp is then used to estimate the variable importance, which is printed and plotted.

Sample Dataset

```
set.seed(7)
basep <- base[sample(1:6340852,1000),-112]
positivos <- base[base$class == 1,-112]
basep <- rbind(positivos,basep)
```

Excluding the class variable

```
basepclass <- basep$class
basep$class <- NULL
```

Converting the variables to numeric

```
basep[] <- lapply (basep, function (x) as.numeric (as.character (x)))
```

Adding the class variable

```
basepclass -> basep$class
```

Factorizing the class variable

```
basep$class <- as.factor(basep$class)
```

Wrangling for NA´s

```
basep <- na.omit(basep)
```

Several variables with zero variances will be removed from the Dataset

```
basep$MODFOINT <-NULL
basep$CPP <- NULL
basep$CHECK <-NULL
basep$MODEMAILOTR <- NULL
basep$EFECTI <- NULL
basep$MODDIRCOM <- NULL
basep$MODDIROTR <- NULL
basep$MODFONCOM <- NULL
basep$MODFONINT <- NULL
basep$RESCLASEGD <- NULL
basep$MAILPAGPENS <- NULL
basep$CLADINCONF <-  NULL
basep$CLADINMOD1 <- NULL
basep$CLADINMOD2 <-  NULL
basep$WVPA <- NULL
basep$SOLRETCCV <- NULL
```

Prepare the training scheme

```
control <- trainControl(method="repeatedcv", number=10, repeats=3)
```

Train the model

```
model <- train(class~., data=basep, method="lvq", preProcess="scale", trControl=control)
```

Estimate variable importance

```
importance <- varImp(model, scale=FALSE)
```

Summarize importance

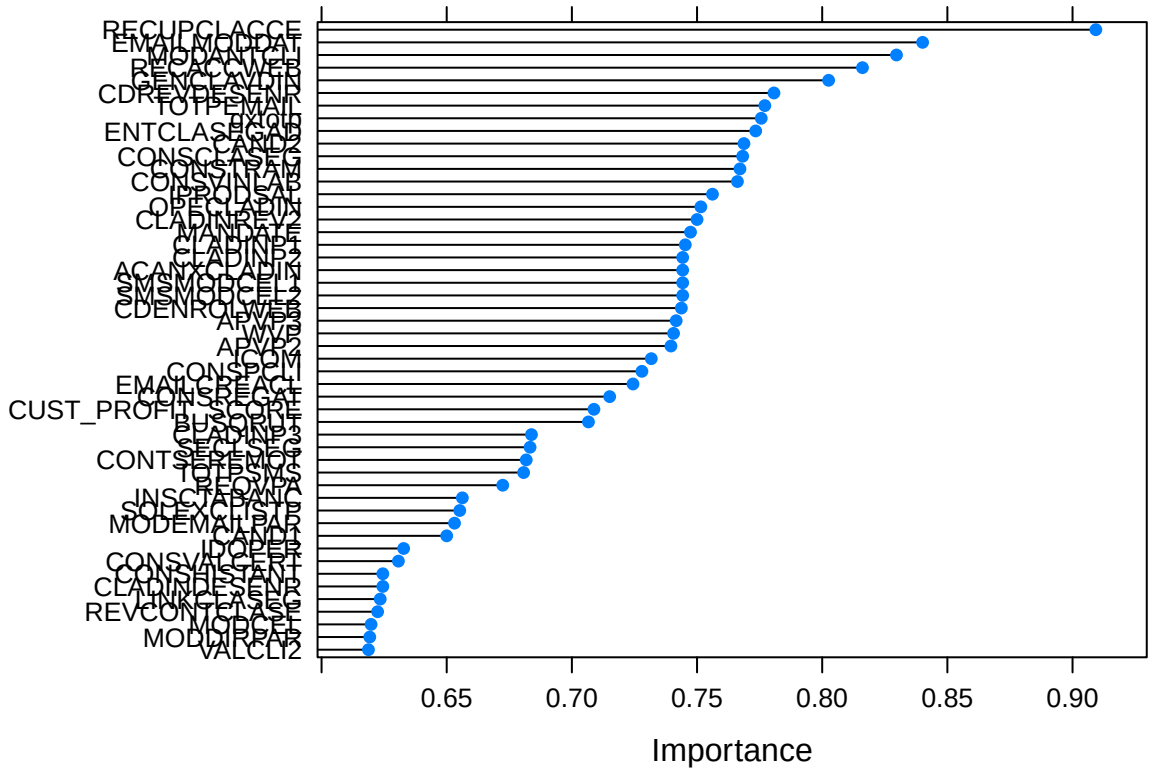
```
print(importance, top = 50)
```

```
## ROC curve variable importance
##
##   only 50 most important variables shown (out of 98)
##
##                   Importance
```

```
## RECUPCLACCE          0.9093
## EMAILMODDAT          0.8401
## MODANTCLI            0.8297
## RECACCWEB            0.8161
## GENCLAVDIN           0.8026
## CDREVDESENR          0.7807
## TOTPEMAIL            0.7771
## qxtotp               0.7756
## ENTCLASEGAD          0.7734
## CAND2                0.7687
## CONSCLASEG           0.7682
## CONSTRAM             0.7671
## CONSVINLAB           0.7661
## IPRODSAL             0.7562
## OPECLADIN            0.7515
## CLADINREV2           0.7500
## MANDATE              0.7474
## CLADINP1             0.7453
## CLADINP2             0.7443
## SMSMODCEL1           0.7443
## SMSMODCEL2           0.7443
## ACANXCLADIN          0.7443
## CDENROLWEB           0.7437
## APVP3                0.7417
## WVP                  0.7406
## APVP2                0.7396
## ICOM                 0.7317
## CONSPCLI             0.7280
## EMAILCREACL          0.7244
## CONSREGAT            0.7151
## CUST_PROFIT_SCORE    0.7088
## BUSQRUT              0.7066
## CLADINP3             0.6839
## SECLSEG              0.6833
## CONTSEREMOT          0.6818
## TOTPSMS              0.6807
## REQVPA               0.6724
## INSCTABANC           0.6562
## SOLEXCLISTP          0.6552
## MODEMAILPAR          0.6531
## CAND1                0.6500
## IDOPER               0.6328
## CONSVALCERT          0.6307
## CLADINDESENR         0.6245
## CONSHISTANT          0.6245
## LINKCLASEG           0.6234
## REVCONTCLASE         0.6224
## MODCEL               0.6198
## MODDIRPAR            0.6193
## VALCLI2              0.6187
```

Plot importance

```
plot(importance, top = 50)
```



There is a manual check as there are variables with high correlation but due to business requirements, those need to be included in the data set.

Based on the results on both variable redundace methods and business requirements, a new data set will be created.

```
rm(basep)
```

New Data Set with relevant variables, based on variable redundace methods applied

```
base <- base[,c('ACANXCLADIN',
                'APVP2',
                'APVP3',
                'AVCETRAM',
                'CAND2',
                'CAVP2AUT',
                'CDMODCEL',
                'CDREVDESENR',
                'CLADINCONF',
                'CLADINDESENR',
                'CLADINMOD1',
                'CLADINMOD2',
                'CLADINP1',
```

```
                'CLADINREV1',
                'CLADINREV2',
                'CONSCLASEG',
                'CONSTRAM',
                'CONSVINLAB',
                'DEPOSIT',
                'CUSTOMER_AGE',
                'EMAILMODDAT',
                'ENTCLASEGAD',
                'GENCLAVDIN',
                'IPRODSAL',
                'MANDATE',
                'MCLAACCFOR',
                'MCLASATFOR',
                'MODANTCLI',
                'MODEMAILCOM',
                'OPECLADIN',
                'qxtotp',
                'RECACCWEB',
                'RECUPCLACCE',
                'REPAVTRAM',
                'BALANCE',
                'CUST_PROFIT_SCORE',
                'CUST_SERVICE_SCORE',
                'SECLSEG',
                'SMSMODCEL1',
                'SMSMODCEL2',
                'REQVP',
                'REQVPA',
                'WVP',
                'TOTPEMAIL',
                'dominio',
                'class')]
```

Delete records without selected transactions

```
base$borrar <- ifelse(
  base$ACANXCLADIN == 0 &
    base$APVP2 == 0 &
    base$APVP3 == 0 &
    base$AVCETRAM == 0 &
    base$CAND2 == 0 &
    base$CAVP2AUT == 0 &
    base$CDMODCEL == 0 &
    base$CDREVDESENR == 0 &
    base$CLADINCONF == 0 &
    base$CLADINDESENR == 0 &
    base$CLADINMOD1 == 0 &
    base$CLADINMOD2 == 0 &
    base$CLADINP1 == 0 &
    base$CLADINREV1 == 0 &
    base$CLADINREV2 == 0 &
    base$CONSCLASEG == 0 &
```

```
    base$CONSTRAM == 0 &
    base$CONSVINLAB == 0 &
    base$DEPOSIT == 0 &
    #base$edad_cliente == 0 &
    base$EMAILMODDAT == 0 &
    base$ENTCLASEGAD == 0 &
    base$GENCLAVDIN == 0 &
    base$IPRODSAL == 0 &
    base$MANDATE == 0 &
    base$MCLAACCFOR == 0 &
    base$MCLASATFOR == 0 &
    base$MODANTCLI == 0 &
    base$MODEMAILCOM == 0 &
    base$OPECLADIN == 0 &
    #base$qxtotp == 0 &
    base$RECACCWEB == 0 &
    base$RECUPCLACCE == 0 &
    base$REPAVTRAM == 0 &
    base$BALANCE == 0 &
    #base$score_rentabilidad == 0 &
    #base$score_servicio == 0 &
    base$SECLSEG == 0 &
    base$SMSMODCEL1 == 0 &
    base$SMSMODCEL2 == 0 &
    base$REQVP == 0 &
    base$REQVPA == 0 &
    base$WVP == 0 &
    base$TOTPEMAIL == 0 ,1,0)

table(base$borrar)
```

```
##
##       0       1
## 4995214 1685989
```

Base with selected variables and transactions with movements

```
base <- base[base$borrar == 0,]
```

Clean up (NA´s) and save the dataset

```
base <- na.omit(base)

table(base$class)
```

```
##
##       0       1
## 4645956      16
```

Additional wrangling to include risky web domain (business requirement)

```
base$dominioriesgoso <- ifelse(base$dominio == 'vtr.net' | base$dominio == 'mi.cl',1,0)
base$dominioriesgoso <- as.factor(base$dominioriesgoso)

table(base$dominioriesgoso)
```

```
##
##       0       1
## 4576263   69709
```

```
base$dominio <- NULL
```

1.3.2) DATA SET ANALYSIS

#Additional analysis over the new Data set (base)

```
cat("\nBase set dimension :",dim(base))
```

```
##
## Base set dimension : 4645972 47
```

```
cat("\nNumber of unique ages :",base$CUSTOMER_AGE %>% unique() %>% length())
```

```
##
## Number of unique ages : 108
```

```
cat("\nNumber of unique profitability score  :",base$CUST_PROFIT_SCORE %>% unique() %>% length())
```

```
##
## Number of unique profitability score  : 47691
```

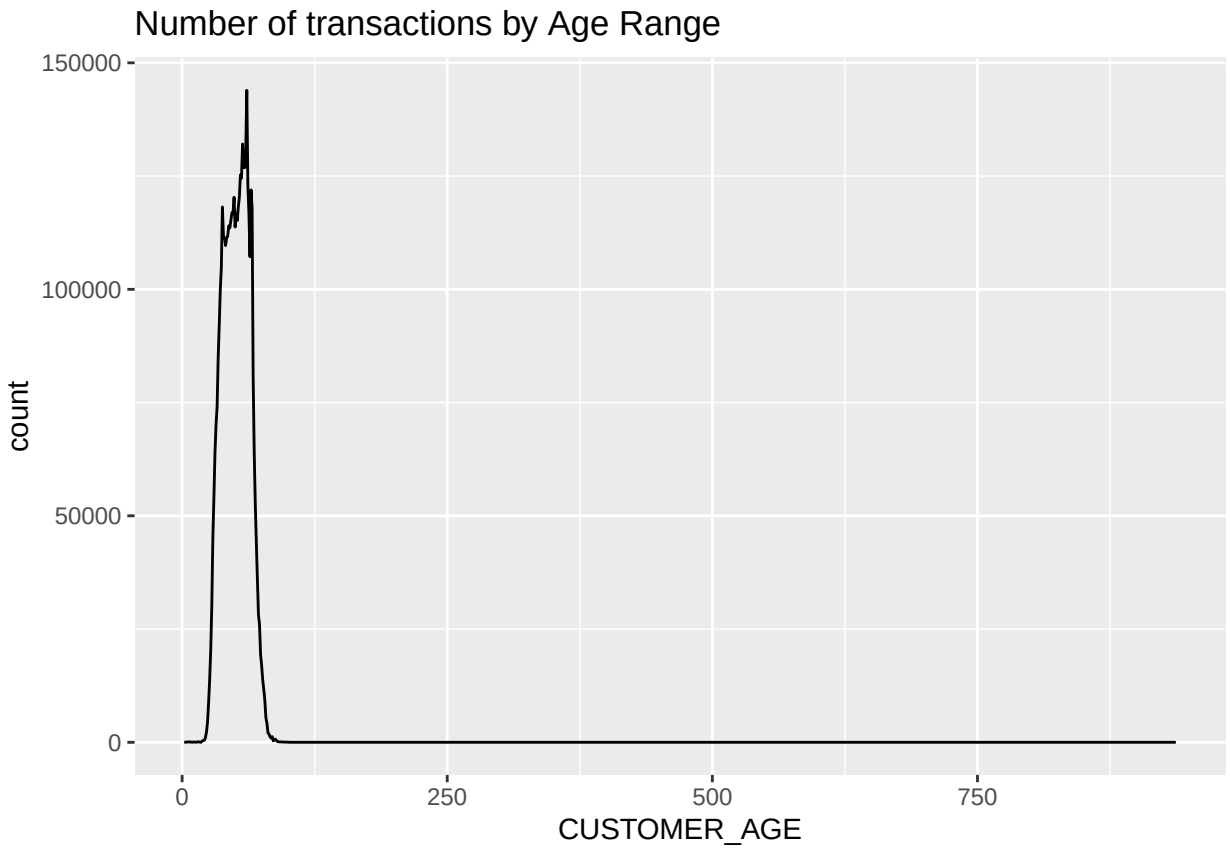
```
cat("\nNumber of unique service score  :",base$CUST_SERVICE_SCORE %>% unique() %>% length())
```

```
##
## Number of unique service score  : 227
```

#Number of transactions by Age Range

```
base %>%
  group_by(CUSTOMER_AGE) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = CUSTOMER_AGE, y = count)) +
  geom_line() +
  ggtitle("Number of transactions by Age Range")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

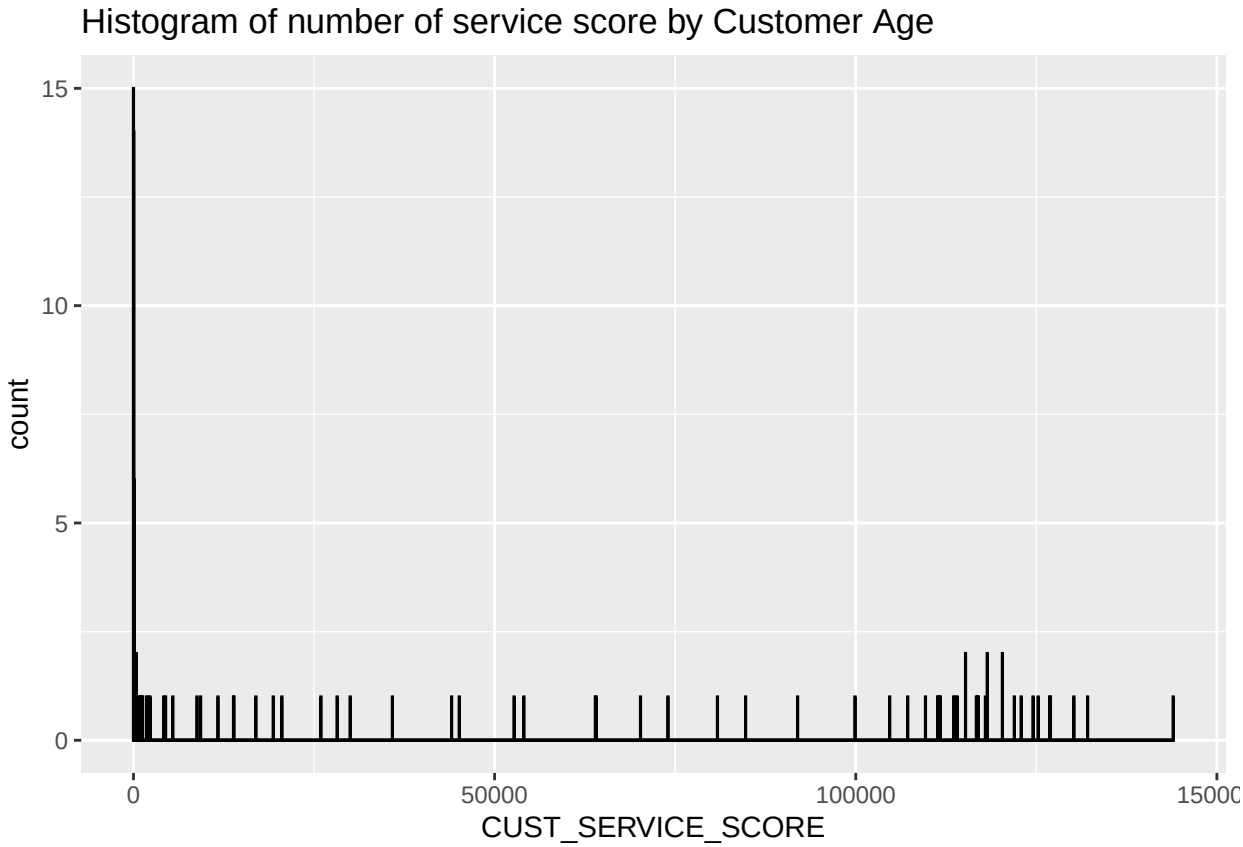


#Histogram of number of transactions for each service score

```
base %>%
  group_by(CUSTOMER_AGE) %>%
  summarise(CUST_SERVICE_SCORE=n()) %>%
  ggplot(aes(CUST_SERVICE_SCORE)) +
  geom_histogram(color="black", binwidth = 50) +

  ggtitle("Histogram of number of service score by Customer Age")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



#CLASS variable Analysis - Variable used to identify negative == 0 & positive == 1 fraud transactions. Only 16 transactions with suspicious (fraud) activity

```
table(base$class)
```

```
##
##       0       1
## 4645956      16
```

2) METHODS/ANALYSIS

2.1) DATA WRANGLING

Split for training and test; training with 80% of the initial data set

```
set.seed(123)
v <- c(1:(nrow(base)*1))
variables <- c(4:ncol(base))
train_test_split <- initial_split(base[v,variables], prop = 0.80)
train_test_split
```

```
## <Analysis/Assess/Total>
## <3716778/929194/4645972>
```

Functions training() and testing() used to create train and test data sets

```
train_tbl <- training(train_test_split)
test_tbl  <- testing(train_test_split)

nrow(train_tbl)
```

```
## [1] 3716778
```

```
nrow(test_tbl)
```

```
## [1] 929194
```

Train Data set : 3.716.778 records Test Data set : 929.194 records

```
table(train_tbl$class)
```

```
##
##       0       1
## 3716765      13
```

```
table(test_tbl$class)
```

```
##
##      0      1
## 929191      3
```

Suspicious transaction in the train data set : 13 Suspicious transaction in the test data set : 3 ——————————————-

Split for validation Data Set. 50% of the test data set will be used for validation

```
set.seed(123)
porcvalidac <- nrow(test_tbl) * 0.5
filasaletorias <- sample(1:nrow(test_tbl),porcvalidac)
tbl_validacion <- test_tbl[filasaletorias,]
table(tbl_validacion$class)
```

```
##
##      0      1
## 464596      1
```

```
test_tbl <- test_tbl[-filasaletorias,]
```

Positive transactions (fraud) proportion in Data Sets

```
table(train_tbl$class)
```

```
##
##       0       1
## 3716765      13
```

```
table(test_tbl$class)
```

```
##
##      0      1
## 464595      2
```

```
table(tbl_validacion$class)
```

```
##
##      0      1
## 464596      1
```

Suspicious transaction in the train data set : 13 Suspicious transaction in the test data set : 2 Suspicious transaction in the validation data set : 1

Based on the fraud proportions, it is clear we have a data sampling issue that needs to be addressed using data balancing techniques

2.2) DATA BALANCE

Data balance technique must be applied as the variable used (class) to identify suspicious transactions is not equally distributed. This will create a challenge for the training process as it will be difficult to identify logical rules.

To train the models we should have 20% on suspicious (positive) and 80% on negative transactions. A new training Data Set will be created.

2.2.1) Undersampling - Decrease negative (not suspicious) transactions

```
set.seed(123456)
# Positive cases for training
qx <- 13
qxn <- qx * 4
# Training data set assembly
negativos <- train_tbl[train_tbl$class == 0, ]
tbl_negativos <- negativos[sample(1:nrow(negativos), qxn),]
tbl_positivos <- train_tbl[train_tbl$class == 1, ]

train_tbl_manual <- rbind(tbl_negativos, tbl_positivos)

# Checking for the new Training data set
nrow(train_tbl_manual)
```

```
## [1] 65
```

```
head(train_tbl_manual)
```

```
##         AVCETRAM CAND2 CAVP2AUT CDMODCEL CDREVDESENR CLADINCONF CLADINDESENR
## 5055031        0     0        0        0           0          0            0
## 5339770        0     0        0        0           0          0            0
## 4432455        0     0        0        0           0          0            0
## 1088257        0     0        0        0           0          0            0
## 326371         0     0        0        0           0          0            0
## 5422580        0     0        0        0           0          0            0
##         CLADINMOD1 CLADINMOD2 CLADINP1 CLADINREV1 CLADINREV2 CONSCLASEG
## 5055031          0          0        0          0          0          0
## 5339770          0          0        0          0          0          0
## 4432455          0          0        0          0          0          0
## 1088257          0          0        0          0          0          0
## 326371           0          0        0          0          0          0
## 5422580          0          0        0          0          0          0
##         CONSTRAM CONSVINLAB DEPOSIT CUSTOMER_AGE EMAILMODDAT ENTCLASEGAD
## 5055031        0          0       0           66           0           0
## 5339770        0          0       0           64           0           0
## 4432455        0          0       0           67           0           0
## 1088257        0          0       0           47           0           0
## 326371         0          0       0           54           0           0
## 5422580        0          0       0           62           0           0
##         GENCLAVDIN IPRODSAL MANDATE MCLAACCFOR MCLASATFOR MODANTCLI MODEMAILCOM
## 5055031          0        1       0          0          0         0           0
## 5339770          0        0       0          0          0         0           0
## 4432455          0        0       0          0          0         0           0
## 1088257          0        0       0          0          0         0           0
## 326371           0        0       0          0          0         0           0
## 5422580          0        0       0          0          0         0           0
##         OPECLADIN qxtotp RECACCWEB RECUPCLACCE REPAVTRAM BALANCE
## 5055031         0      0         0           0         0       0
## 5339770         0      0         0           0         0       1
## 4432455         0      0         0           0         0       1
## 1088257         0      1         1           1         0       1
## 326371          0      0         0           0         0       1
## 5422580         0      0         0           0         0       1
##         CUST_PROFIT_SCORE CUST_SERVICE_SCORE SECLSEG SMSMODCEL1 SMSMODCEL2
## 5055031             21500               1080       0          0          0
## 5339770             -2905               1115       0          0          0
## 4432455              9249               1270       0          0          0
## 1088257             -2905               1165       1          0          0
## 326371              25455               1125       0          0          0
## 5422580              2266               1235       0          0          0
##         REQVP REQVPA WVP TOTPEMAIL class borrar dominioriesgoso
## 5055031     0      0   0         0     0      0               0
## 5339770     0      0   0         0     0      0               0
## 4432455     0      0   0         0     0      0               0
## 1088257     0      0   0         1     0      0               0
## 326371      0      0   0         0     0      0               0
## 5422580     0      0   0         0     0      0               0
```

```
table(train_tbl_manual$class)
```

```
##
##  0  1
```

```
## 52 13
```

Number of negative (not fraud) transactions : 52 Number of positive (fraud) transactions : 13

2.2.2) Oversampling - Increase positive (suspicious) transactions

```
positivos <- train_tbl[train_tbl$class == 1, ]

# Increasing positives

n <- 5

for(i in 1:(n-1)) {

  positivos <- rbind(positivos, positivos)

}

negativos <- train_tbl[train_tbl$class == 0, ]
indnegativos <- sample(1:nrow(negativos), (nrow(positivos)*4))
tbl_negativos <- negativos[indnegativos,]
train_tbl_manual <- rbind(tbl_negativos, positivos)

table(train_tbl_manual$class)
```

```
##
##   0   1
## 832 208
```

Number of negative (not fraud) transactions : 832 Number of positive (fraud) transactions : 208 ————————

Saving the data sets

```
save(train_tbl_manual,file="train_tbl_manual.RData")
save(tbl_validacion,file="validacion_tbl.RData")
save(train_tbl,file="train_tb_completa.RData")
save(test_tbl,file="test_tb_completa.RData")

# Deleting objects to release memory
rm(list = ls())
```

2.3) VARIABLE TREATMENT AND DATA CLEANING

```
load("train_tbl_manual.RData")
load("test_tb_completa.RData")
train_tbl_manual$DELETE <- NULL
test_tbl$DELETE <- NULL
```

Factorizing the class variable (target variable to train the algorithm)

```
train_tbl_manual$class <- as.factor(train_tbl_manual$class)
test_tbl$class <- as.factor(test_tbl$class)
```

Cleaning the training data set

```
p <- as.data.frame(summary(train_tbl_manual))
p <- na.omit(p)
```

Additional wrangling for special transactions (deprecated transactions based on business definition)

```
p1 <- sqldf("select Var2 as q from p where Freq not like '%1: 0%' group by Var2 having count(Var2) > 1")
```

Wrangling - Removing blankspaces from the names and adding to the data frame

```
p1$q <- gsub(pattern = "\\s",
             replacement = "",
             x = p1$q)


incluir <- p1$q

train_tbl_manual <- train_tbl_manual[,incluir]
```

Moving the target/class variable to the end of the table

```
target<- train_tbl_manual$class
train_tbl_manual$class <- NULL
target -> train_tbl_manual$class
```

Excluding the target variable (class) from the test data set for prediction

```
x<-test_tbl[,-42]
```

Wrangling -excluding NA´s from the train data set

```
train_tbl_manual<- na.omit(train_tbl_manual)
```

2.4) DATA MODELING - MACHINE LEARNING ALGORITHMS

Training with several Machine Learning Models

2.4.1) NAIVE BAYES ALGORITHM

Naive Bayes is a Supervised Machine Learning algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach. It is based on the idea that the predictor variables in a Machine Learning model are independent of each other. Meaning that the outcome of a model depends on a set of independent variables that have nothing to do with each other.

Build the model

```
modelBayes<-naiveBayes(class~.,data=train_tbl_manual)
```

Summarize the model

```
summary(modelBayes)
```

```
##           Length Class  Mode
## apriori    2     table  numeric
## tables    43     -none- list
## levels     2     -none- character
## isnumeric 43     -none- logical
## call       4     -none- call
```

Predict using the model

```
test_tbl$pred_Bayes<-predict(modelBayes,x)
```

Accuracy of the model

```
mtab1<-table(test_tbl$pred_Bayes,test_tbl$class, dnn = c("prediccion", "real"))
confusionMatrix(mtab1, positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           real
## prediccion      0      1
##          0 455811      0
##          1   8784      2
##
##                Accuracy : 0.9811
##                  95% CI : (0.9807, 0.9815)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 4e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.000e+00
##             Specificity : 9.811e-01
##          Pos Pred Value : 2.276e-04
##          Neg Pred Value : 1.000e+00
##              Prevalence : 4.305e-06
##          Detection Rate : 4.305e-06
##    Detection Prevalence : 1.891e-02
##       Balanced Accuracy : 9.905e-01
##
##        'Positive' Class : 1
##
```

Saving model´s accuracy

```
cm1<- confusionMatrix(mtab1, positive = '1')
overall.accuracy1<-cm1$overall['Accuracy']
```

Saving the model

```
save(modelBayes, file = "modelBayes.rda")
```

2.4.2) RANDOM FOREST ALGORITHM

Random forest algorithm is a supervised classification and regression algorithm. As the name suggests, this algorithm randomly creates a forest with several trees.

Generally, the more trees in the forest the more robust the forest looks like. Similarly, in the random forest classifier, the higher the number of trees in the forest, greater is the accuracy of the results.

In simple words, Random forest builds multiple decision trees (called the forest) and glues them together to get a more accurate and stable prediction. The forest it builds is a collection of Decision Trees, trained with the bagging method.

Build the model

```
model15<-randomForest(class ~ ., data=train_tbl_manual[,-1], ntree=600)
```

Summarize the model

```
summary(model15)
```

```
##                 Length Class  Mode
## call               4   -none- call
## type               1   -none- character
## predicted       1040   factor numeric
## err.rate        1800   -none- numeric
## confusion          6   -none- numeric
## votes           2080   matrix numeric
## oob.times       1040   -none- numeric
## classes            2   -none- character
## importance        42   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            14   -none- list
## y               1040   factor numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

Predict using the model

```
test_tbl$pred_randomforest<-predict(model15,x)
```

Accuracy of the model

```
mtab2<-table(test_tbl$pred_randomforest,test_tbl$class, dnn = c("prediction", "real"))
confusionMatrix(mtab2, positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           real
## prediction      0      1
##          0 464539      0
##          1     56      2
##
##                Accuracy : 0.9999
##                  95% CI : (0.9998, 0.9999)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0667
##
##  Mcnemar's Test P-Value : 1.987e-13
##
##             Sensitivity : 1.000e+00
##             Specificity : 9.999e-01
##          Pos Pred Value : 3.448e-02
##          Neg Pred Value : 1.000e+00
##              Prevalence : 4.305e-06
##          Detection Rate : 4.305e-06
##    Detection Prevalence : 1.248e-04
##       Balanced Accuracy : 9.999e-01
##
##        'Positive' Class : 1
##
```

Saving model´s accuracy

```
cm2<- confusionMatrix(mtab2, positive = '1')
overall.accuracy2<-cm2$overall['Accuracy']
```

Saving the model

```
save(model15, file = "model15_RF.rda")
```

2.4.3) KNN ALGORITHM

KNN which stand for K Nearest Neighbor is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points.

Build the model

```
model9<-knn3(class ~ .,data=train_tbl_manual,k=14)
```

Summarize the model

```
summary(model9)
```

```
##         Length Class  Mode
## learn    2     -none- list
## k        1     -none- numeric
## terms    3     terms  call
## xlevels 38     -none- list
## theDots  0     -none- list
```

Predict using the model

```
test_tbl$pred_knn<-predict(model9,x,type="class")
```

Accuracy of the model

```
mtab3<-table(test_tbl$pred_knn,test_tbl$class, dnn = c("prediccion", "real"))
confusionMatrix(mtab3, positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           real
## prediccion      0      1
##          0 396036      0
##          1  68559      2
##
##                Accuracy : 0.8524
##                  95% CI : (0.8514, 0.8535)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.000e+00
##             Specificity : 8.524e-01
##          Pos Pred Value : 2.917e-05
##          Neg Pred Value : 1.000e+00
##              Prevalence : 4.305e-06
##          Detection Rate : 4.305e-06
##    Detection Prevalence : 1.476e-01
##       Balanced Accuracy : 9.262e-01
##
##        'Positive' Class : 1
##
```

Saving model´s accuracy

```
cm3<- confusionMatrix(mtab3, positive = '1')
overall.accuracy3<-cm3$overall['Accuracy']
```

Saving the model

```
save(model9, file = "modeloknn2020.rda")
```

3) RESULTS

3.1) MACHINE LEARNING MODEL ACCURACY

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified.

The accuracy will be used as the variable to select the algorithm to be used for validation (and eventually for production purposes)

3.2) MACHINE LEARNING MODEL VALIDATION

```
MODEL_EVALUATED<- c("Bayes Model", "RF Model", "KNN Model")
MODEL_ACCURACY<- c(overall.accuracy1, overall.accuracy2, overall.accuracy3)
EVALUATION_RESULT<- data.frame(MODEL_EVALUATED, MODEL_ACCURACY)
EVALUATION_RESULT
```

```
##   MODEL_EVALUATED MODEL_ACCURACY
## 1     Bayes Model      0.9810933
## 2        RF Model      0.9998795
## 3       KNN Model      0.8524334
```

Based on the results processing over training and test data sets, the Random forest Algorithm is providing the best accuracy. The RF algorithm will be used to process against the validation data set.

Naive Bayes could be considered as a second alternative as it´s accuracy is close to the RF.

KNN accuracy is out of the accuracy range we are looking for; further recommendations will be provided in the CONCLUSION section on this report.

3.3) MACHINE LEARNING - SELECTED MODEL EXECUTION AGAINST VALIDATION DATA SET

Loading validation data set

```
load("validacion_tbl.RData")
```

Checking for fraud (positive == 1) transactions

```
table(tbl_validacion$class)
```

```
##
##      0      1
## 464596      1
```

Excluding the target variable (class) from the validation data set for prediction

```
x_final<-tbl_validacion[,-42]
```

Predict using the model

```
tbl_validacion$pred_randomforest<-predict(model15,x_final)
```

Accuracy of the model

```
mtabfinal<-table(tbl_validacion$pred_randomforest,tbl_validacion$class, dnn = c("prediccion", "real"))
confusionMatrix(mtabfinal, positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           real
## prediccion      0      1
##          0 464545      0
##          1     51      1
##
##                Accuracy : 0.9999
##                  95% CI : (0.9999, 0.9999)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0377
##
##  Mcnemar's Test P-Value : 2.534e-12
##
##             Sensitivity : 1.000e+00
##             Specificity : 9.999e-01
##          Pos Pred Value : 1.923e-02
##          Neg Pred Value : 1.000e+00
##              Prevalence : 2.152e-06
##          Detection Rate : 2.152e-06
##    Detection Prevalence : 1.119e-04
##       Balanced Accuracy : 9.999e-01
##
##        'Positive' Class : 1
##
```

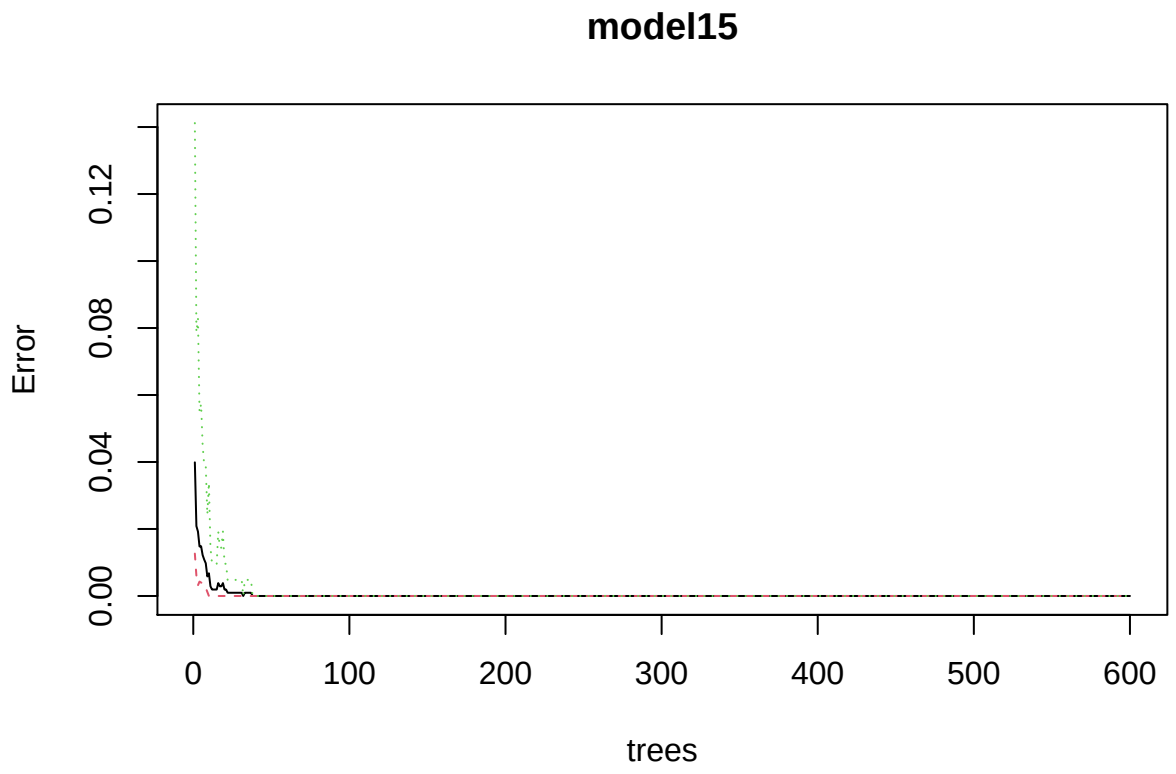
Getting model´s accuracy

```
cmfinal<- confusionMatrix(mtabfinal, positive = '1')
overall.accuracyfinal<-cmfinal$overall['Accuracy']
overall.accuracyfinal
```

```
##  Accuracy
## 0.9998902
```

Plotting the model

```
plot(model15)
```



3.3) MACHINE LEARNING EXECUTION - OBSERVATION

After 50 iterations (trees), real vs prediction trend to have the same values. Accuracy : 0.9999053

4) CONCLUSION

4.1) INSIGHTS

The accuracy obtained with the RF algorithm, will provide efficiencies to the bank /financial institution as they will avoid manual checkings once a complaint with a possible fraud is received. The Bank will save money as suspicious / fraud transactions will be held until further validation it´s done with the customer limiting Bank´s exposure to unnecesary reputation and regulation´s risks.

The Machine Learning platform and the results has proven to be an effective method for (predictive) fraud detection.

4.2) RECOMMENDATIONS

- Increase data set volume for training and test purposes.
- Include additional techniques to improve training and testing processes (i.e Cross validation)

4.3) NEXT STEPS

- Automate the training and model/algorithm selection process, based on its accuracy. (AutoML)
- Increase data processing capacity using a platform like DATABRICKS.
- Deploy a real time application using the trained model; the application should be calling a Rest API passing the transaction to validate as a parameter.
- Further investigation on API´s deployment (Shiny vs Plumber) must be performed.