The Analysis Model for Default of Credit Card Clients Data Set

CAGIN KESKIN [1]

[1]*Applied Statistics, Graduate School, Izmir University of Economics*
Teleferik Mahallesi, Sakarya Cd. No:156, 35330 Balçova/İzmir, Republic of Turkey

1. Introduction

The data set is defined that a Taiwan-based credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. This would inform the issuer's decisions on who to give a credit card to and what credit limit to provide. It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers. The credit card issuer has gathered information on 30000 customers. Dataset contains information on 24 variables, including demographic factors, credit data, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not? Data set information is as follows; *Y* : Customers who have default payment (Yes = 1 , No = 0) *X1* : Amount of the given credit ; it includes both the individual consumer credit and her/his family *X2* : Gender (1 = Male , 2 = Female) *X3* : Education (1 = Graduate school, 2 = University, 3 = High School, 4 = Others) *X4* : Marital Status (1 = Married, 2 = Single, 3 = Others) *X5* : Age (Year) *X6* - *X11* : History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: *X6* = the repayment status in September 2005; *X7* = the repayment status in August 2005; …; *X11* = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; …; 8 = payment delay for eight months; 9 = payment delay for nine months and above.*X12* - *X17*: Amount of bill statement (NT dollar). *X12* = amount of bill statement in September 2005; *X13* = amount of bill statement in August 2005; …; *X17* = amount of bill statement in April 2005 *X18* - *X23*: Amount of previous payment (NT dollar). *X18* = amount paid in September 2005; *X19* = amount paid in August 2005; …; *X23* = amount paid in April 2005.

1.1 Describe Data Set

In this study aim is to detect that are which variables effected results of decision process and which of them is strongest? According to the information given, the describe of the data set is as follows; Numbers are used to characterize the expressions in the sex, education, marriage and history of past payment (repayment status) but the numbers do not have a numerical significance. Therefore, Y, X2, X3, X4, X6-X11 included in the categorical dataset. The expressions in the limit balance, age, amount of bill statement and amount of previous payment are numerical values, so X1, X5, X12-X17, X18-X23 in numerical data set. They are divided two groups which are discrete (X5) and continuous (X1, X12-X17, X18-X23). For measurement levels are divided quantitative and qualitative dataset. In quantitative data occurred ratio data (X1, X12-X17, X18-X23) and interval data (X5). In qualitative data occurred ordinal data (X3, X6-X11) and nominal data (Y, X2, X4).

Required Packages:

```r
library(openxlsx)

library(ggplot2)

library(C50)

library(funModeling)

library(gridExtra)

library(MASS)
library(corrplot)

library(nnet)

library(caret)

library(neuralnet)

library(dplyr)

library(e1071)
```

Importing Default Credit Cards Data

```r
data.path = "C:/Users/asus-pc/Desktop/Proje_R/default\ of\ credit\ card\ c
lients/CreditCardsFixed.xlsx"
raw.data  = read.xlsx(data.path,sheet = 1)

raw.data$SEX = as.factor(raw.data$SEX)
levels(raw.data$SEX) = c("Male","Female")
```

Converting sex data to factor.

```r
raw.data$EDUCATION = as.factor(raw.data$EDUCATION)
levels(raw.data$EDUCATION) = c("Unknown","Graduate school", "University",
                                "High school", "Others","Unknown", "Unknow)
```

Converting education levels to factor.

```r
raw.data$MARRIAGE = as.factor(raw.data$MARRIAGE)
levels(raw.data$MARRIAGE) = c("Unknown","Married","Single","Others")
```

Converting marrige levels to factor.

```r
raw.data$PAY_0 = as.factor(raw.data$PAY_0)
raw.data$PAY_2 = as.factor(raw.data$PAY_2)
raw.data$PAY_3 = as.factor(raw.data$PAY_3)
raw.data$PAY_4 = as.factor(raw.data$PAY_4)
raw.data$PAY_5 = as.factor(raw.data$PAY_5)
raw.data$PAY_6 = as.factor(raw.data$PAY_6)
```

Converting repatment status to factor.

```r
raw.data$default.payment.next.month = as.factor(raw.data$default.payment.n
ext.month)
```

```r
levels(raw.data$default.payment.next.month) = c("No","Yes")
colnames(raw.data)[colnames(raw.data) == "default.payment.next.month"] = "
PAID"

Converting default payment to factor.

credits = raw.data
```

Raw data pre-processing complete let's rename it. Here we provide an overview of the data set we have using the *summary* function. We will then examine this information with the help of graphs. We have not NA values however, some values are unknow so, we cleared data set and we obtain new data set that is approximately 4000 variables left.

```r
credits <- credits[(credits$EDUCATION=="Graduate school" | credits$EDUCATI
ON=="University" | credits$EDUCATION=="High school"),]
credits$EDUCATION <- factor(credits$EDUCATION)

credits <- credits[(credits$MARRIAGE=="Married" | credits$MARRIAGE=="Singl
e"),]
credits$MARRIAGE <- factor(credits$MARRIAGE)
credits <- credits[(credits$PAY_0!=-2 & credits$PAY_0!=0),]
credits <- credits[(credits$PAY_2!=-2 & credits$PAY_2!=0),]
credits <- credits[(credits$PAY_3!=-2 & credits$PAY_3!=0),]
credits <- credits[(credits$PAY_4!=-2 & credits$PAY_4!=0),]
credits <- credits[(credits$PAY_5!=-2 & credits$PAY_5!=0),]
credits <- credits[(credits$PAY_6!=-2 & credits$PAY_6!=0),]

summary(credits[2:length(colnames(credits))])
```

```
##    LIMIT_BAL          SEX                   EDUCATION       MARRIAGE
##  Min.   : 10000   Male  :1634   Graduate school:1674   Married:2071
##  1st Qu.: 70000   Female:2352   University     :1698   Single :1915
##  Median :150000                 High school    : 614
##  Mean   :172143
##  3rd Qu.:240000
##  Max.   :740000
##
##       AGE             PAY_0          PAY_2          PAY_3
##  Min.   :21.00   -1     :2347   -1     :2417   -1     :2407
##  1st Qu.:29.00   2      : 784   2      :1325   2      :1336
##  Median :35.00   1      : 624   3      : 139   3      : 121
##  Mean   :36.47   3      : 158   4      :  58   4      :  49
##  3rd Qu.:43.00   4      :  34   7      :  20   7      :  26
##  Max.   :72.00   8      :  19   6      :  12   6      :  23
##                  (Other):  20   (Other):  15   (Other):  24
##      PAY_4          PAY_5          PAY_6         BILL_AMT1
##  -1     :2493   -1     :2538   -1     :2504   Min.   : -4316.0
##  2      :1230   2      :1169   2      :1245   1st Qu.:   931.5
##  3      : 104   3      : 130   3      : 125   Median :  4373.0
##  4      :  62   4      :  75   7      :  45   Mean   : 22138.5
##  7      :  56   7      :  55   4      :  40   3rd Qu.: 22205.5
##  5      :  34   5      :  14   6      :  14   Max.   :581775.0
```

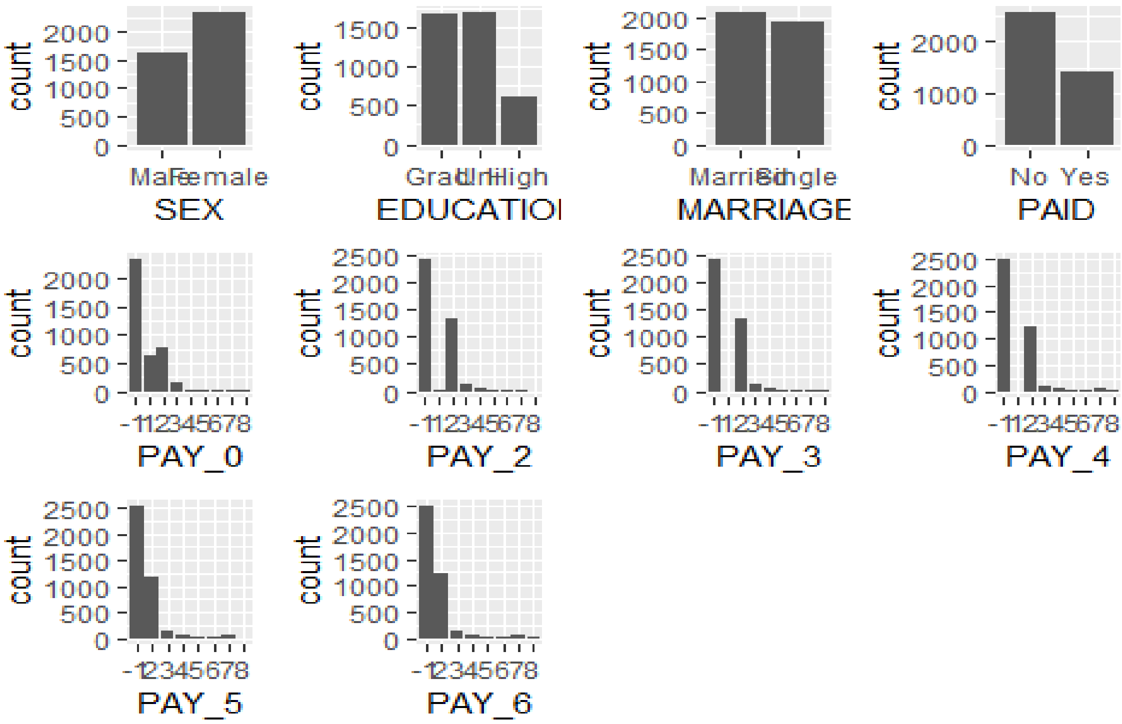
```
##  (Other):   7   (Other):   5   (Other):  13
##    BILL_AMT2        BILL_AMT3        BILL_AMT4         BILL_AMT5
##  Min.    :-24704   Min.    :-61506   Min.    : -3903.0   Min.    : -3876
##  1st Qu.:   856   1st Qu.:   836   1st Qu.:   833.8   1st Qu.:   846
##  Median :  4398   Median :  4218   Median :  4186.0   Median :  4082
```

```
##  Mean    : 22349   Mean    : 22365   Mean    : 22674.7   Mean    : 22670
##  3rd Qu.: 22817   3rd Qu.: 23012   3rd Qu.: 22848.8   3rd Qu.: 23287
##  Max.    :572677   Max.    :471175   Max.    :486776.0   Max.    :503914
##
##    BILL_AMT6         PAY_AMT1         PAY_AMT2         PAY_AMT3
##  Min.    :-339603   Min.    :     0   Min.    :     0   Min.    :     0
##  1st Qu.:     780   1st Qu.:   316   1st Qu.:   316   1st Qu.:   316
##  Median :    4162   Median :  1600   Median :  1600   Median :  1443
##  Mean    :  22783   Mean    :  4673   Mean    :  4580   Mean    :  4704
##  3rd Qu.:  23999   3rd Qu.:  4408   3rd Qu.:  4400   3rd Qu.:  4200
##  Max.    : 527711   Max.    :187206   Max.    :302961   Max.    :417588
##
##     PAY_AMT4          PAY_AMT5          PAY_AMT6        PAID
##  Min.    :     0.0   Min.    :     0.0   Min.    :     0   No :2567
##  1st Qu.:   326.2   1st Qu.:   109.5   1st Qu.:     0   Yes:1419
##  Median :  1443.5   Median :  1240.0   Median :  1046
##  Mean    :  4550.2   Mean    :  4613.3   Mean    :  4605
##  3rd Qu.:  4100.0   3rd Qu.:  4004.5   3rd Qu.:  3718
##  Max.    :193712.0   Max.    :303512.0   Max.    :345293
##
```

According to new data set, firstly found the frequency tables of the data set.

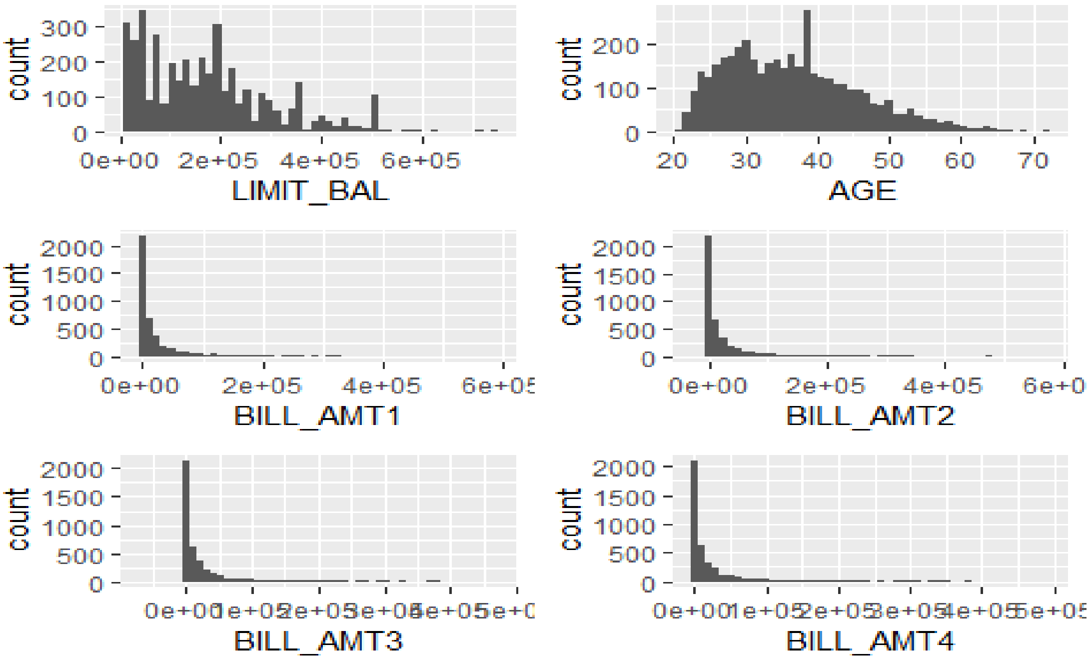
For categorical data variables ;

```
c0 = ggplot(credits, aes(x=SEX)) + geom_bar()
c1 = ggplot(credits, aes(x=EDUCATION)) + geom_bar() + scale_x_discrete(lab
els = c('Grad.','Uni','High'))
c2 = ggplot(credits, aes(x=MARRIAGE)) + geom_bar()
c3 = ggplot(credits, aes(x=PAID)) + geom_bar()
c4 = ggplot(credits, aes(x=PAY_0)) + geom_bar()
c5 = ggplot(credits, aes(x=PAY_2)) + geom_bar()
c6 = ggplot(credits, aes(x=PAY_3)) + geom_bar()
c7 = ggplot(credits, aes(x=PAY_4)) + geom_bar()
c8 = ggplot(credits, aes(x=PAY_5)) + geom_bar()
c9 = ggplot(credits, aes(x=PAY_6)) + geom_bar()
grid.arrange(c0,c1,c2,c3,c4,c5,c6,c7,c8,c9, ncol=4, nrow=3)
```

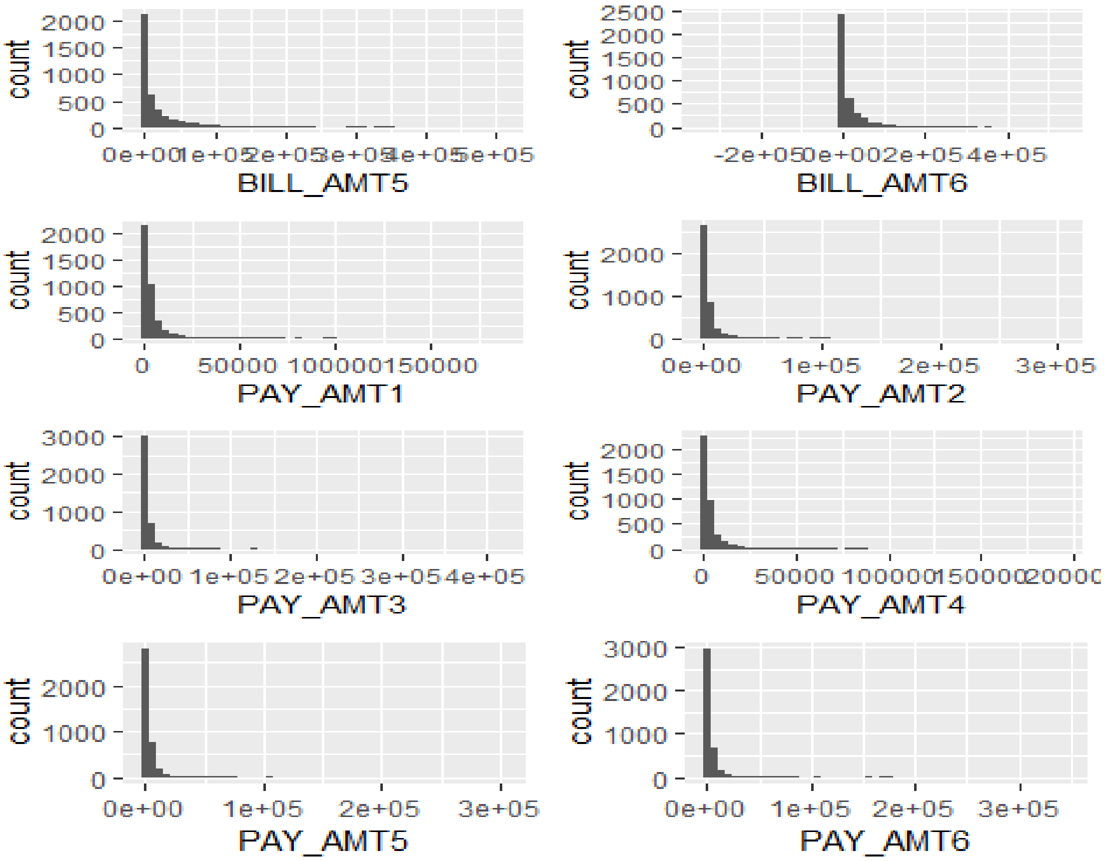


For numerical data variables;

```
binsize = 50
c0 = ggplot(credits, aes(x=LIMIT_BAL)) + geom_histogram(bins=binsize)
c1 = ggplot(credits, aes(x=AGE)) + geom_histogram(bins=binsize)
c2 = ggplot(credits, aes(x=BILL_AMT1)) + geom_histogram(bins=binsize)
c3 = ggplot(credits, aes(x=BILL_AMT2)) + geom_histogram(bins=binsize)
c4 = ggplot(credits, aes(x=BILL_AMT3)) + geom_histogram(bins=binsize)
c5 = ggplot(credits, aes(x=BILL_AMT4)) + geom_histogram(bins=binsize)
c6 = ggplot(credits, aes(x=BILL_AMT5)) + geom_histogram(bins=binsize)
c7 = ggplot(credits, aes(x=BILL_AMT6)) + geom_histogram(bins=binsize)
c8 = ggplot(credits, aes(x=PAY_AMT1)) + geom_histogram(bins=binsize)
c9 = ggplot(credits, aes(x=PAY_AMT2)) + geom_histogram(bins=binsize)
c10 = ggplot(credits, aes(x=PAY_AMT3)) + geom_histogram(bins=binsize)
c11 = ggplot(credits, aes(x=PAY_AMT4)) + geom_histogram(bins=binsize)
c12 = ggplot(credits, aes(x=PAY_AMT5)) + geom_histogram(bins=binsize)
c13 = ggplot(credits, aes(x=PAY_AMT6)) + geom_histogram(bins=binsize)
grid.arrange(c0,c1,c2,c3,c4,c5, ncol=2, nrow=3)
```

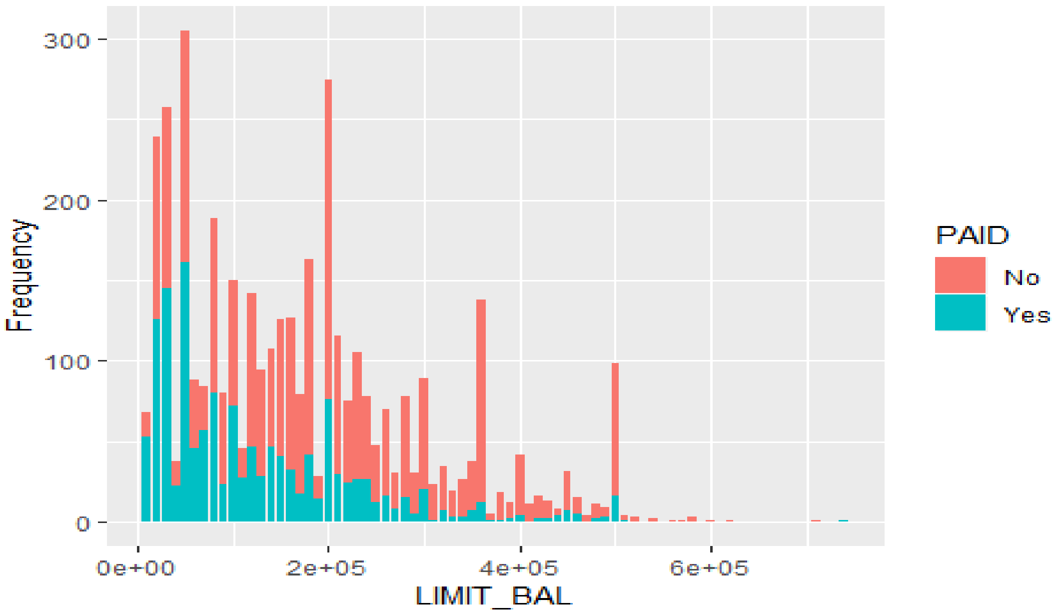



```
grid.arrange(c6,c7,c8,c9,c10,c11,ncol=2,nrow=3)
```



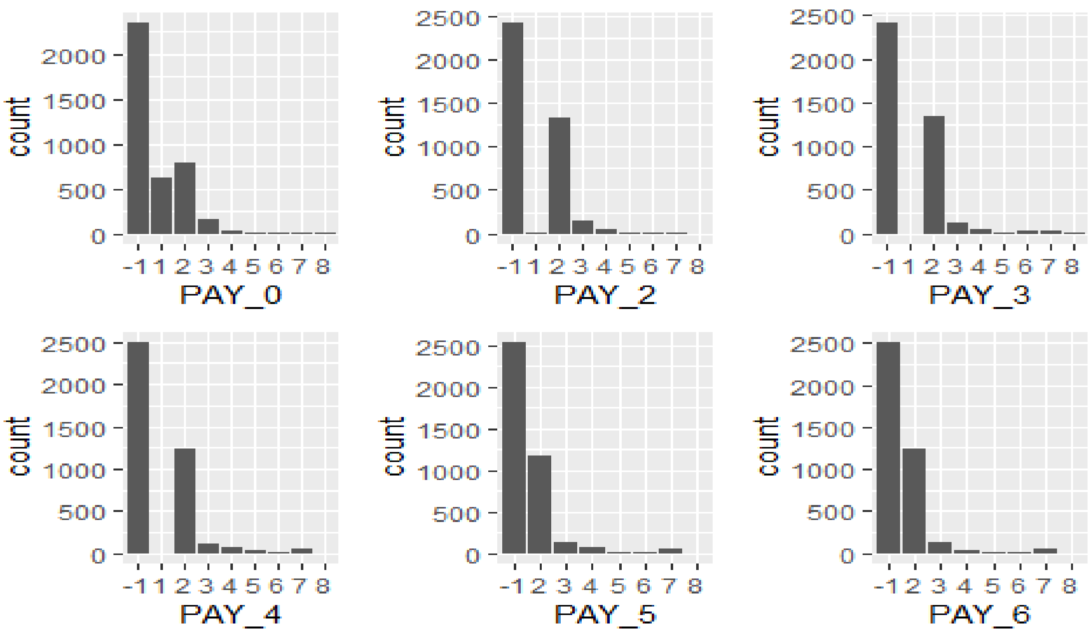

```
grid.arrange(c12,c13,ncol=2,nrow=3)
```

```r
ggplot(data = credits, aes(x = LIMIT_BAL, fill = PAID)) + geom_bar() + yla
b("Frequency")
```



Graph 1. Bar-plot which plot frequency of limit balance levels with respect to paid credit default.

```r
p0 = ggplot(data = credits,aes(x=PAY_0)) + geom_bar()
p2 = ggplot(data = credits,aes(x=PAY_2)) + geom_bar()
p3 = ggplot(data = credits,aes(x=PAY_3)) + geom_bar()
p4 = ggplot(data = credits,aes(x=PAY_4)) + geom_bar()
p5 = ggplot(data = credits,aes(x=PAY_5)) + geom_bar()
p6 = ggplot(data = credits,aes(x=PAY_6)) + geom_bar()
grid.arrange(p0,p2,p3,p4,p5,p6,ncol=3,nrow=2)
```

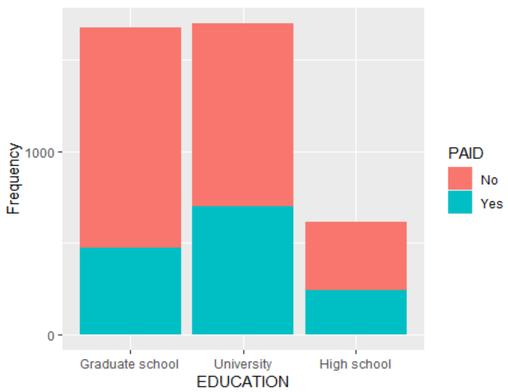


Graph 2. The repayment status in mounths 2005

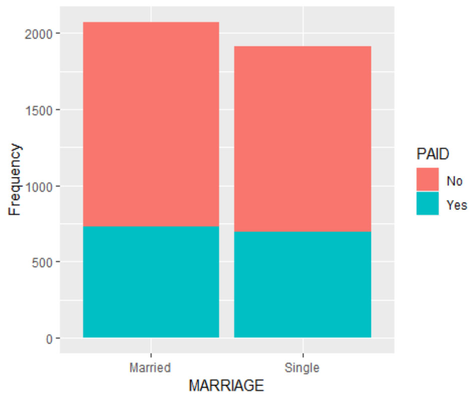
Correlation between data set variables

According to the customer's social status information for next month payments and non-payment frequency graphics are as follows. These graphs allow us to interpret the effects of social situations on payments.First, the graphs of the effects of categorical values on the payment status are given. We then demonstrated the categorical values of 2 with the help of a boxplot, and then we performed a chi-squared test with all categorical data.

```
ggplot(data = credits, aes(x = EDUCATION, fill = PAID)) + geom_bar() +
scale_y_continuous(breaks = seq(min(0),max(30000),by=1000),na.value = T) +
ylab("Frequency")

ggplot(data = credits,aes(x = MARRIAGE , fill = PAID)) + geom_bar() + ylab
("Frequency")
```



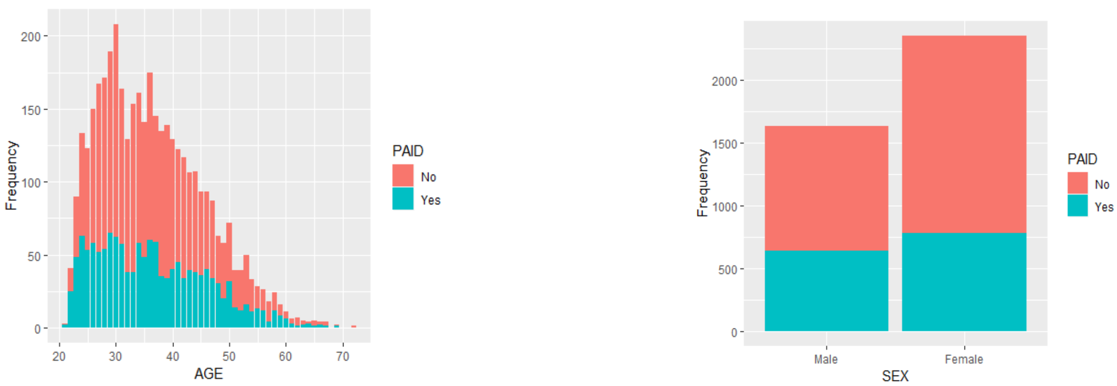
Graph 3. Bar-plot which plot education levels of credit card users



Graph 4. Bar-plot which plot marriage levels of credit card users

After plotting Education, now we create a new bar-plot which plots martial status of credit card users. Now, by creating a histogram chart, we will look at the gender-based age distribution of credit card users.

```
ggplot(data = credits, aes(x = AGE , fill = PAID)) + geom_bar() + ylab("Fr
equency")
```



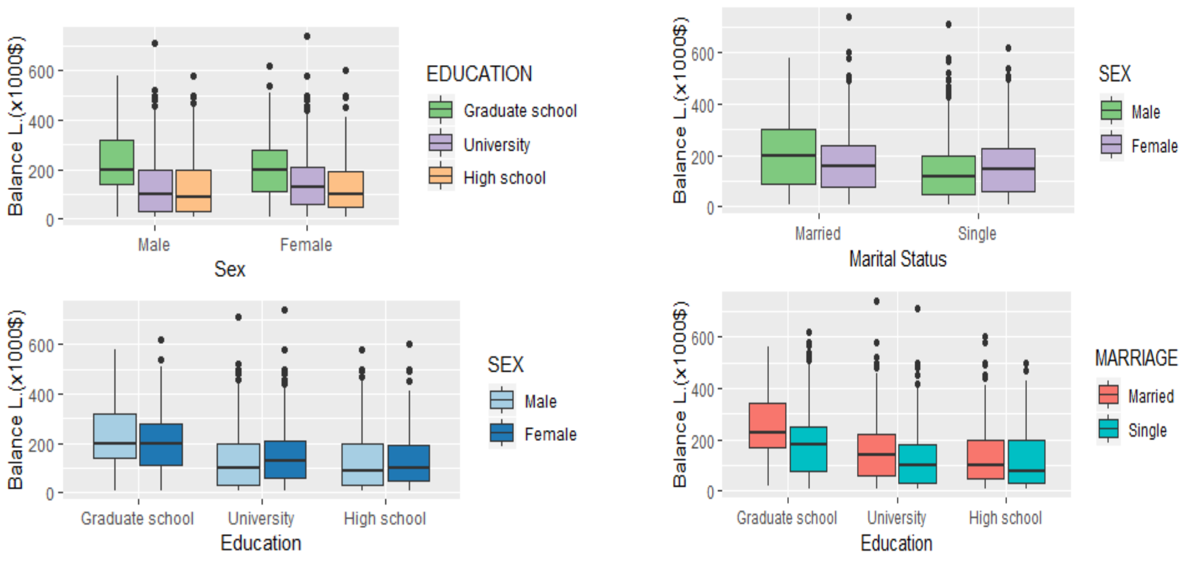
Graph 5. Bar-plot which plot age and sex levels of credit card users

With a similar approach, we will look at the education-based age distribution of credit card users.

```
ggplot(data = credits, aes(x = SEX , fill = PAID)) + geom_bar() + ylab("Fr
equency")
```

We see that there are more women in the data set, but the pay percentage of women is lower than men. To better understand the impact of the three-social status on payments, we will examine a boxplot chart together.

```
bx1 = ggplot(data = credits, aes(x = SEX, y = (LIMIT_BAL/1000), fill=EDUCA
TION)) +
  geom_boxplot() +
  xlab("Sex") +
  ylab("Balance L.(x1000$)") +
  scale_fill_brewer(palette = "Accent")
bx2 = ggplot(credits, aes(x = EDUCATION,y = (LIMIT_BAL/1000), fill=SEX)) +
  geom_boxplot() +
  xlab("Education") +
  ylab("Balance L.(x1000$)") +
  scale_fill_brewer(palette = "Paired")
bx3 = ggplot(data = credits, aes(x = MARRIAGE, y = (LIMIT_BAL/1000), fill=
SEX)) +
  geom_boxplot() +
  xlab("Marital Status") +
  ylab("Balance L.(x1000$)") +
  scale_fill_brewer(palette = "Accent")
bx4 = ggplot(credits, aes(x = EDUCATION, y = (LIMIT_BAL/1000), fill=MARRIA
GE)) +
  geom_boxplot() +
  xlab("Education") +
  ylab("Balance L.(x1000$)")
grid.arrange(bx1,bx2,nrow=2,ncol=1)

grid.arrange(bx3,bx4,nrow=2,ncol=1)
```



Graph 6. Box-plot which compare relations between social status

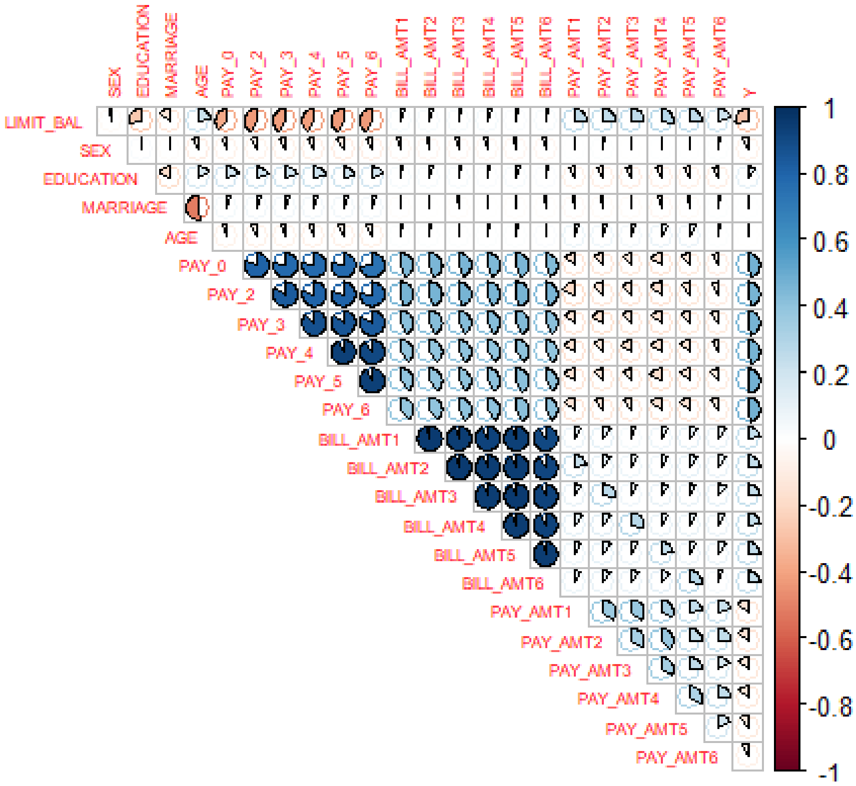
When we compared the balance limits with gender, education and marriage status. We obtained result that gender has no effects on balance limit decision process of bank while the education level is a positive effect on the process. Additionally, we compared sex with respect to marital status and we obtained similar result only female, so can say that there is no change at females' side such as balance limits depending on their marital status. On the other hand, balance limit changes a lot of things side of males with the expenditures which is the reason on increased balance limits. And result of fourth graph is education level is affected on marital status, but marital status is not important on decision process of balance limit. However, here we have evaluated only 3 social statues. We applied a chi-square test to see all categorical data (repayment status and demographic status) in their relationships.

```
data.path = "C:/Users/asus-pc/Desktop/Proje_R/default\ of\ credit\ card\ c
lients/CreditCardsFixed.xlsx"
raw.data  = read.xlsx(data.path,sheet = 1)
raw.data = raw.data[,2:length(colnames(raw.data))]
raw.data = raw.data[(raw.data$EDUCATION == 1 | raw.data$EDUCATION == 2 | r
aw.data$EDUCATION == 3),]
raw.data = raw.data[(raw.data$MARRIAGE == 1 | raw.data$MARRIAGE == 2),]
raw.data <- raw.data[raw.data$PAY_0 != -2 & raw.data$PAY_0 != 0,]
raw.data <- raw.data[raw.data$PAY_2 != -2 & raw.data$PAY_2 != 0,]
raw.data <- raw.data[raw.data$PAY_3 != -2 & raw.data$PAY_3 != 0,]
raw.data <- raw.data[raw.data$PAY_4 != -2 & raw.data$PAY_4 != 0,]
raw.data <- raw.data[raw.data$PAY_5 != -2 & raw.data$PAY_5 != 0,]
raw.data <- raw.data[raw.data$PAY_6 != -2 & raw.data$PAY_6 != 0,]

categorical.data <- raw.data[,c(2,3,4,6:11,24)]
model<-glm(categorical.data$default.payment.next.month~.,family=binomial(l
ink='logit'),data=categorical.data)
anova(model,test="Chisq")

## Analysis of Deviance Table
## Model: binomial, link: logit
## Response: categorical.data$default.payment.next.month
## Terms added sequentially (first to last)
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         3985     5190.4
## SEX          1    14.80      3984     5175.6 0.0001194 ***
## EDUCATION    1    47.60      3983     5128.0 5.238e-12 ***
## MARRIAGE     1     3.13      3982     5124.9 0.0769678 .
## PAY_0        1   892.30      3981     4232.6 < 2.2e-16 ***
## PAY_2        1    94.25      3980     4138.3 < 2.2e-16 ***
## PAY_3        1    19.41      3979     4118.9 1.056e-05 ***
## PAY_4        1    10.47      3978     4108.4 0.0012101 **
## PAY_5        1    11.02      3977     4097.4 0.0009033 ***
## PAY_6        1    10.70      3976     4086.7 0.0010695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our hypothesis:
H_0 = There is no relationship between categoric variable and paid. H_1= There is some relationship between categoric variable and paid. According to our results, we understand that categorical data show binomial distribution. All categorical values except for marital status are effective on payment status. To achieve better results, we rejected Ho for categories that the sex, education status, pay_0, pay_2, pay_3 and pay_5 classes.

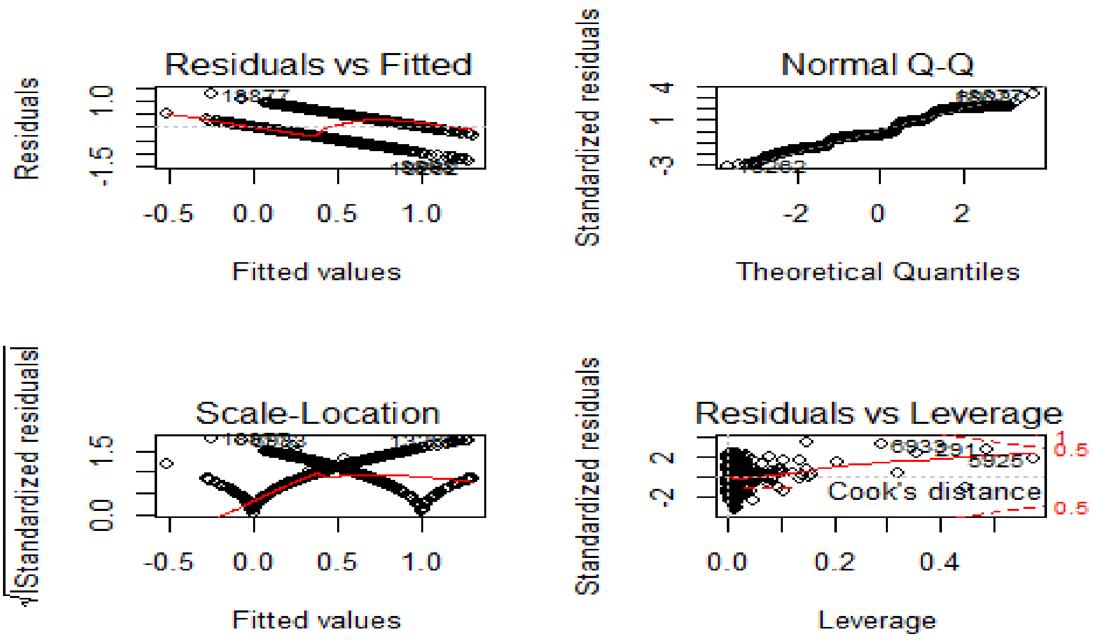
Correlation Table

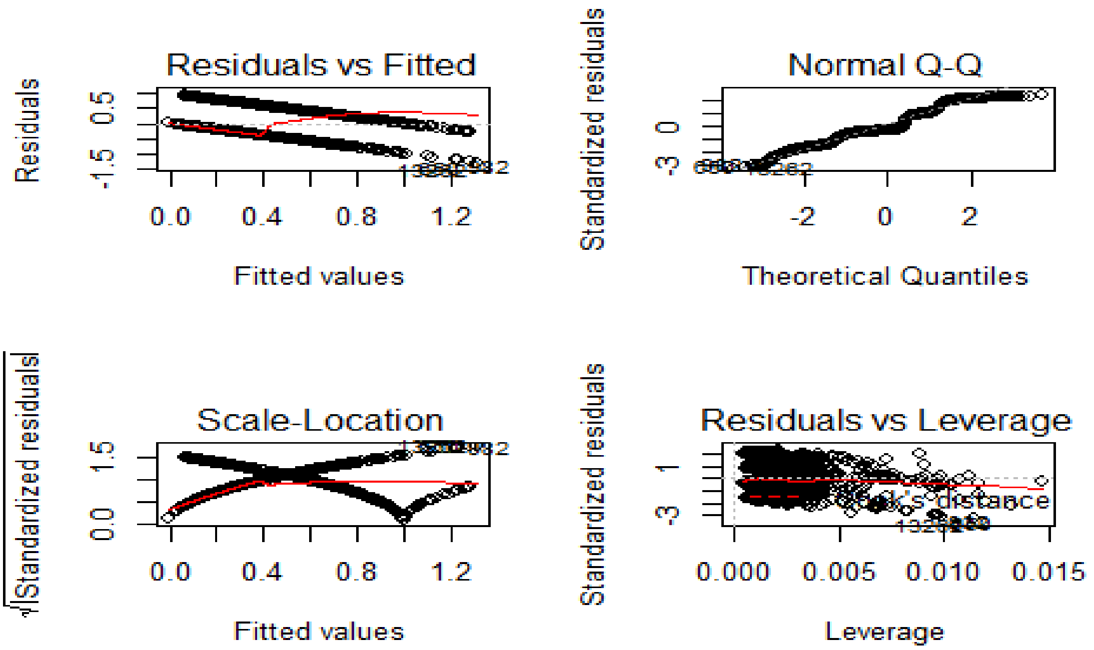
```
names.data = colnames(raw.data)
names.data <- names.data[1:(length(names.data)-1)]
names.data <- c(names.data, "Y")

numeric.cor<-cor(raw.data)
colnames(numeric.cor) <- names.data
rownames(numeric.cor) <- names.data
corrplot(numeric.cor, diag = FALSE,
         tl.pos = "td", tl.cex = 0.5, method = "pie", type = "upper")
```



We used the values that have a high relation with applied correlation. According to result of correlation respectively he lowers the amount of given credit limit of the balance owing, the bigger the chances to default. Male persons have more chances to default. The better education the lower chances to default. The better education the lower chances to default. Having a delay, even for 1 month in any of the previous months, increases the chance of default. The smaller the difference between the amount owed on the bill in September and April, the bigger the chances to default. The smaller the payment amount, the bigger the chance of default for general.

```
lrm=lm(data= raw.data,raw.data$default.payment.next.month~.)
par(mfrow=c(2,2))
plot(lrm)
```






```
imp <- raw.data[,c(1,5,6,7,11,24)]
lrm=lm(data= imp,imp$default.payment.next.month~.)
par(mfrow=c(2,2))
plot(lrm)
```

We used linear regression to see the relationship between the values in the whole data set. Although we have chosen the most important data, we have obtained according to the results of the data set output of Adjusted R-squared is 0.2699. Because there is no correlation with linear regression for data not showing normal distribution logistic regression is used.

Model

Now we will check hypothesizes about predictors' impact on the dependent variable made in the correlation relationship. Also, we will try to implement a model which can predict default with a level higher than majority class classifier which means accuracy of models should be higher than 77.8%.

Logistic Regression

H_0 = There is no relationship between variables and PAID. H_1 = There is some relationship between variables and PAID.

```
set.seed(101)
index <- createDataPartition(credits$PAID,
                             p = 0.7,
                             list = F)
trainSet <- credits[index,]
xTrain <- trainSet %>% select(-PAID)
yTrain <- trainSet$PAID

testSet <- credits[-index,]
fiveMetric <- function(...) c(twoClassSummary(...),
                              defaultSummary(...))
ctrl <- trainControl(method = "cv",
                     number = 5,
                     summaryFunction = fiveMetric,
                     classProbs = T,
                     verboseIter = T)

ctrlSMOTE <- trainControl(method = "cv",
                          number = 5,
                          summaryFunction = fiveMetric,
                          classProbs = T,
                          sampling = "smote",
                          verboseIter = T)

set.seed(101)
glm_model <- train(PAID ~ ., data = credits,
                   method = "glmStepAIC",
                   trControl = ctrl,
                   preProcess = c("nzv","BoxCox"),
                   metric = "Accuracy"
                   )

summary(glm_model)
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1265  -0.6688  -0.4994   0.7213   4.0938
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.977e+00  1.065e+00  -2.796  0.00517 **
## LIMIT_BAL               -4.986e-03  1.791e-03  -2.784  0.00537 **
## `EDUCATIONHigh school`  -2.228e-01  1.149e-01  -1.938  0.05257 .
## AGE                      2.023e+00  4.993e-01   4.051 5.10e-05 ***
## `PAY_0-1`               -1.418e+00  2.213e-01  -6.408 1.47e-10 ***
```

```
## PAY_01                    -1.380e+00  1.986e-01  -6.945 3.78e-12 ***
## PAY_02                    -3.959e-01  1.967e-01  -2.012  0.04420 *
## PAY_22                     4.947e-01  1.217e-01   4.066 4.79e-05 ***
## `PAY_4-1`                 -3.718e-01  1.875e-01  -1.983  0.04734 *
## PAY_52                     3.770e-01  1.396e-01   2.700  0.00693 **
## `PAY_6-1`                 -3.510e-01  1.822e-01  -1.927  0.05401 .
## BILL_AMT1                 -2.142e-05  8.986e-06  -2.384  0.01714 *
## BILL_AMT2                  2.403e-05  9.129e-06   2.632  0.00849 **
## PAY_AMT1                  -4.481e-05  1.119e-05  -4.006 6.19e-05 ***
## PAY_AMT2                  -4.792e-05  1.047e-05  -4.576 4.73e-06 ***
## PAY_AMT5                  -6.253e-06  4.578e-06  -1.366  0.17198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 5190.4  on 3985  degrees of freedom
## Residual deviance: 3849.5  on 3970  degrees of freedom
## AIC: 3881.5
## Number of Fisher Scoring iterations: 6
```

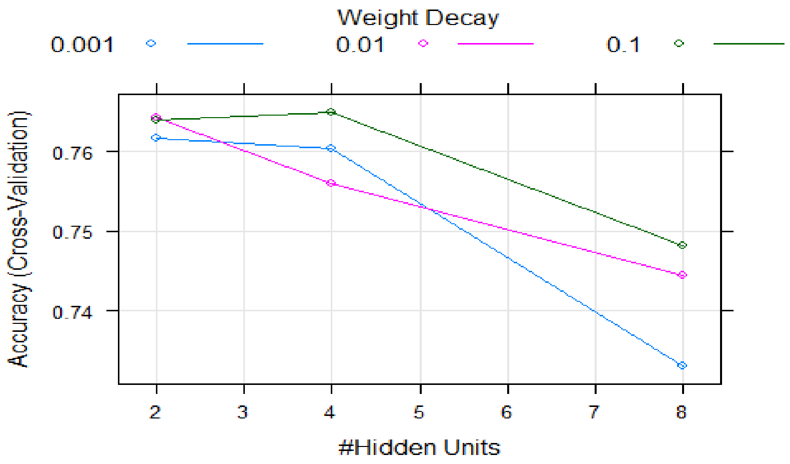
Random Forest Classification

```
set.seed(101)
rf_model <- train(PAID ~ ., data = credits,
                  method = "rf",
                  trControl = ctrl,
                  metric = "Accuracy",
                  tuneGrid = expand.grid(.mtry = c(4,8,12,22)))
## Aggregating results
## Selecting tuning parameters
## Fitting mtry = 8 on full training set

varImp(rf_model)

## rf variable importance
##   only 20 most important variables shown (out of 77)
##
##             Overall
## PAY_5-1     100.00    ## PAY_AMT1    75.57
## BILL_AMT6    92.52    ## PAY_AMT4    75.36
## BILL_AMT4    92.37    ## PAY_AMT2    74.98
## BILL_AMT2    90.97    ## PAY_AMT6    71.58
## BILL_AMT1    88.95    ## PAY_AMT5    68.38
## BILL_AMT5    88.56    ## PAY_6-1     67.02
## BILL_AMT3    86.83    ## PAY_4-1     64.45
## PAY_AMT3     80.85    ## PAY_3-1     62.52
## AGE          79.87
## LIMIT_BAL    78.57
```

Artificial Neural Network

```
nn_grid <- expand.grid(.size = c(4,8,2),
                       .decay = c(0.001,0.01,0.1))
```

```
set.seed(101)
nn_model <- train(x = xTrain, y = yTrain,
                  method = "nnet",
                  tuneGrid = nn_grid,
                  preProcess = c("center", "scale"),
                  trControl = ctrl)

plot(nn_model)
```



Conclusion

The data selected according to the correlation relationship were used logistic regression, random forest and neural network methods respectively. The accuracy rate was 0.78, 0.78 and 0.76, respectively. We would choose Random Forest model because it is still the top choices by combination of other parameters and shows stable result. Additionally, combinations of accuracy, sensivity and specificity for Random Forest is a little bit better than for the Logistic Regression model and Neural Net Work.

References

[1] Ajay Venkatesh, Shomona Gracia Jacob *Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers Volume 145-No.7,July 2016.*

[2] Jian Sun *Analyzing Default Payments of Credit Card Clients in Taiwan* December 2016.

[3] Max Merikoski, Ari Vitala, Nourhan Shafik *Predicting and Preventing Credit Card Default* 2018

[4] Sunakshi Sharma, Vipul Mehra *Default Payment Analysis of Credit Card Clients* July 2018