Statistical modeling and Forecasting analysis of Credit card default

u1801116

May 2021

Introduction

The purpose of this report is to study which customers will default on credit card repayment and which customers will repay normally. To analyze the data, we will establish a logistics regression model and make classification prediction. We will predict credit defaults through the use of customer characteristics. Through customer information characteristics, financial institutions are then able to use provided relevant models and effective information to check for default behaviours, thus reducing financial risks.

The Problem

Credit Risk, also known as default risk, refers to the possibility that the borrower or counterparty is unwilling or unable to perform the contract conditions for various reasons, which constitutes a default, resulting in losses to the bank or counterparty.

This is the main risk in the process of bank credit granting, and it is also a common risk in the process of bank credit card approval. The premise of credit card risk assessment is to obtain the credit information of the applicant, evaluate the obtained information of the applicant, quantify its performance ability, and finally assess whether to grant credit. In the past, credit granting to credit card applicants mainly depended on the evaluation of loan officers, and the evaluation results were influenced by subjective factors, or banks set up special credit decision-making committees to comprehensively evaluate applicants. However, the expansion of consumer loans and micro-loans in recent years has made the previous methods of relying on manual credit evaluation have limitations, and the manual evaluation also has some impersonality and incompleteness.

At present, there are many problems in consumer credit, such as customers are difficult to keep, the loan amount is small, the default rate is high, there is no mortgage guarantee, and adjustment is difficult. It is necessary to cooperate with a more efficient, convenient, accurate, objective and lower cost credit analysis and evaluation scheme. At present, the financial market is facing great uncertainty, and both banks and even all financial institutions and investors are facing the test of huge credit risks. It is an urgent problem for all major financial institutions to establish a sound and effective investor credit evaluation system. Studies have shown that the bank's profit is closely related to the effectiveness of the credit evaluation model, and every 1% increase in the accuracy of the model, the bank can generate billions of returns. In addition, practical research shows that machine learning plays a positive role in the practice of establishing credit risk assessment model, which not only improves its accuracy but also expands its application scope.

This paper tries to find a more accurate credit risk assessment system by applying the research concept of machine learning to the establishment of credit risk assessment model for credit card applicants, so as to help relevant financial institutions establish a more complete customer credit risk assessment system. The credit card risk studied in this paper is mainly to predict whether the trustee will have overdue repayment or non-repayment behavior for more than a large amount of time. The traditional credit card risk assessment is mainly through artificial judgment or discriminant analysis. Nowadays, most of the research uses Logistic regression model. This paper will model customers' credit characteristics and self-information to predict those customers who will repay on time and those who will be in arrears, hoping to provide relevant information for financial institutions to reduce financial risks.

Data Sources

Data is extracted from "default of credit card clients Data Set" donated by I-Cheng Yeh and from Dua, D. and Graff, C. (2019). UCI Machine Learning Repository http://archive.ics.uci.edu/ml Irvine, CA: University of California, School of Information and Computer Science.

The variables used in this report include:

- Credit limit;
- Education level;
- Marital status;
- Gender;
- Age in years;
- The repayment status in date;
- Amount of bill statement in date;
- Amount paid in date;
- Dependent variable: Default payment (0=No,1=Yes)

Data Cleaning

We will first conduct data cleaning before we fit our model for prediction, there are three steps to data cleaning:

1. Missing values treatment
2. Transformations to our variables
3. Checking for correlation between variables

Missing values

Carry out preliminary data cleaning to see if the training and testing data has any missing values and identify the type of missing values. Since the data came from the same source, we are able to combine both sets of data for missing value analysis.

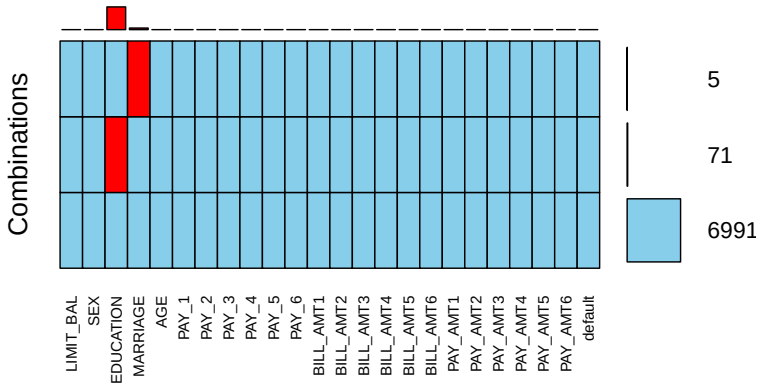

Figure 1: Missing value in each variables

As we can see from the table above, there are 71 missing values on the variable **EDUCATION**, representing educational level, and 5 missing values on the variable **MARRIAGE**, representing marriage status. Since there are only 5 missing values out of 7067 observations for our **MARRIAGE** variable, missing value data might not be sufficient to do much investigation. Therefore, we will only look into our **EDUCATION** variable and take the same approach for both variables.

By using Mosaic plots, we will look into the customer characteristics variables such **LIMIT_BAL**, **AGE** and **SEX** against our missing value variable **EDUCATION**. We only picked customer characteristics variables as these are variables that define individual customers and we do not what specific type of customers to be left out of the model.

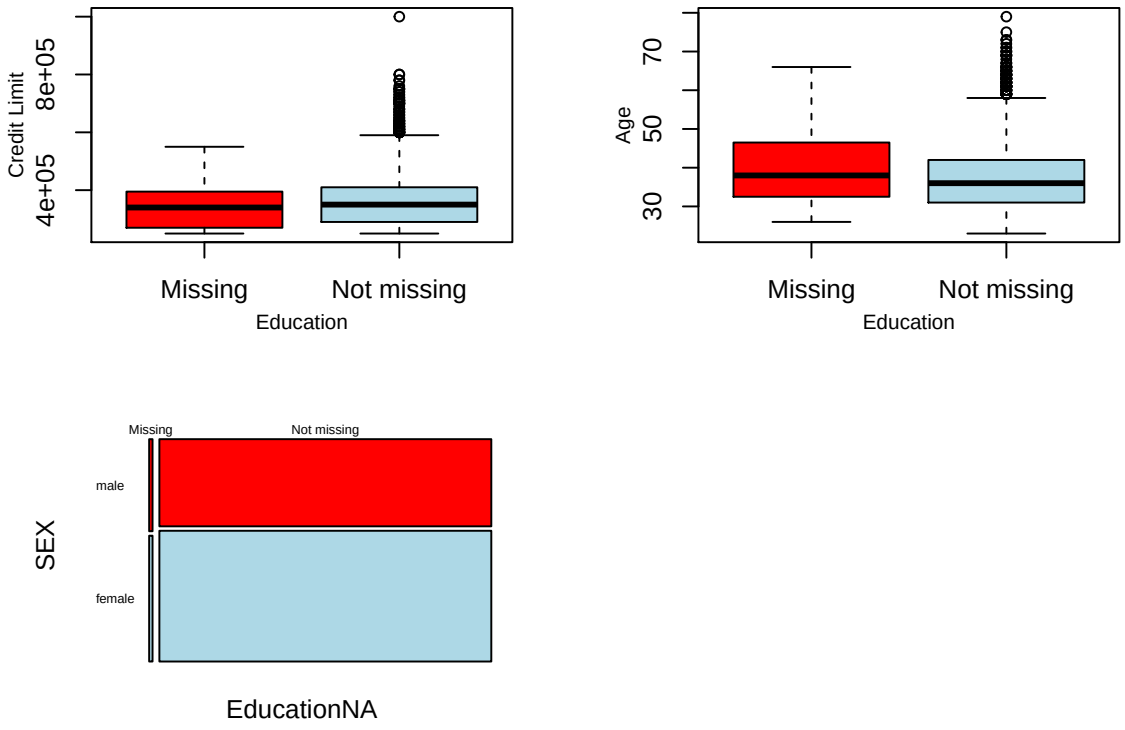

Figure 2: Mosaic Plots for missing value analysis

As we can see from the above Mosaic Plots, there are not much differences in the distribution of our

customer characteristics variables when comparing observations that have missing **EDUCATION** variable and observations that have not missing **EDUCATION** variable. Therefore, we come into a conclusion that these missing values are Missing Completely at Random (MCAR) and we will just delete them from the our training and testing data sets.

Transformations

When performing transformations, we will only transform our variables in our training data. As our dependent variable **default** is a binary variable, we will not do any transformations towards it. Looking into our continuous variables, as **BILL_AMT** and **PAY_AMT** variables are split into many sub variables but are very similar, we will only look into one of each.

After looking into our continuous variable distributions, we conclude that the following transformations would be appropriate:

- Log2 Transformations on **LIMIT_BAL** (Skewness reduced from 1.05 to 0.49)
- Log2+1 Transformations on **PAY_AMT** (Skewness for **PAY_AMT1** reduced from 8.94 to -1.30)

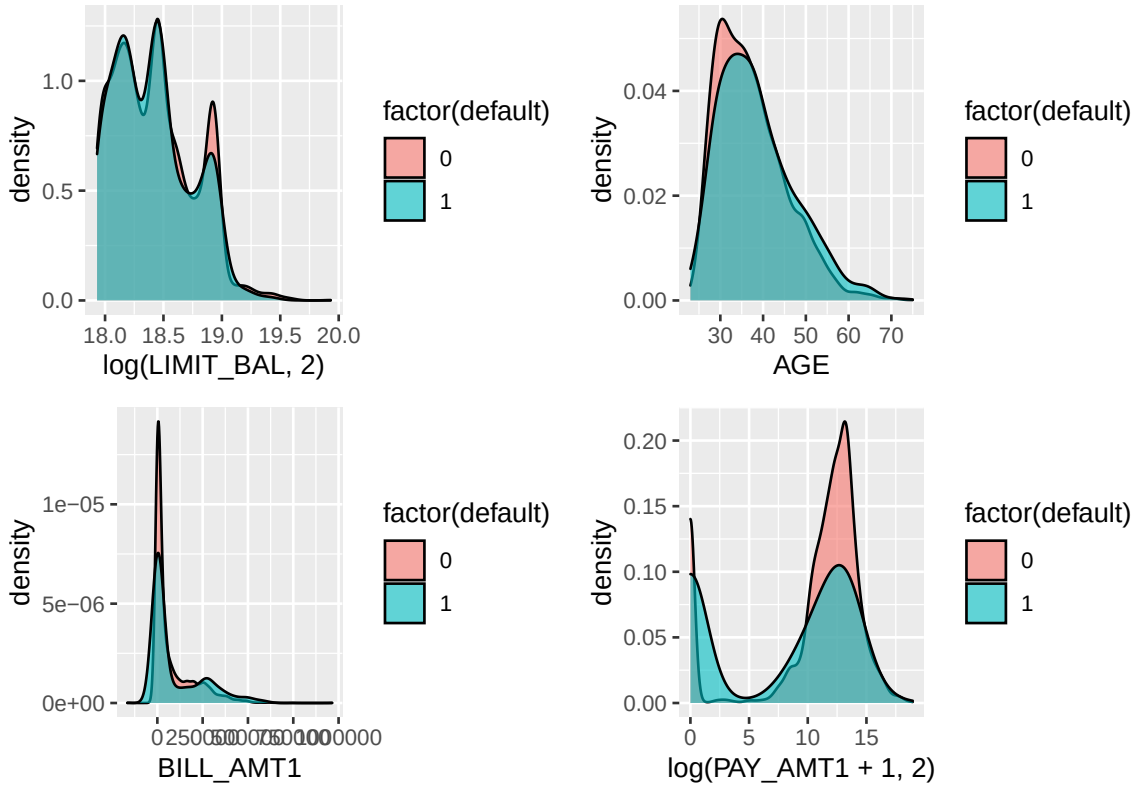

Figure 3: Density Plots of continuous variables after suggested transformations

Looking into our factor variables, there were too many layers in our variable **PAY**, which represents repayment status, and these made the variable hard to interpret. Hence, we decided to conduct the following transformation to our **PAY** variable:

- -2,-1,0 becomes 0 which represents repaid
- 1 remains 1 which represents payment delay for one month
- 2 remains 2 which represents payment delay for two months
- 3,4,5,6,7,8,9 becomes 3 which represents payment delay for three months and above

After doing the above simple transformations, we did not make any further changes to our predictor variables as we still wanted to maintain stronger explanatory power for our model but over transforming our predictor variable will make our model harder to interpret.

Correlation

For continuous variables, we will use Variance Inflation Factor(VIF) to check for correlations. Bear in mind that we are not using VIF as a method of variable selection, variable selection will be something that we will be doing when building our model. In this case, we are only diagnosing multicollinearity between our predictor variables and gathering appropriate conclusions. With variables that are highly correlated and explains the same meanings, we will not include several of these variables in our model as it merely increase the complexitiy of our model but does no good.

Table 1: VIF before and after variable selection

Variable	Coefficient
LIMIT_BAL	1.0372
AGE	1.0179
BILL_AMT1	11.3850
BILL_AMT2	17.9410
BILL_AMT3	13.7380
BILL_AMT4	14.4900
BILL_AMT5	15.6800
BILL_AMT6	9.4740
PAY_AMT1	1.8369
PAY_AMT2	2.1053
PAY_AMT3	2.0975
PAY_AMT4	2.0724
PAY_AMT5	2.0673
PAY_AMT6	1.8510

Variable	Coefficient
LIMIT_BAL	1.0359
AGE	1.0171
BILL_AMT1	2.9482
BILL_AMT6	3.1329
PAY_AMT1	1.7280
PAY_AMT2	1.9270
PAY_AMT3	1.9659
PAY_AMT4	1.9445
PAY_AMT5	1.9664
PAY_AMT6	1.8455

After removing the variables **BILL_AMT2**, **BILL_AMT3**, **BILL_AMT4**, **BILL_AMT5**, we can see that all variable have a VIF less than 5, which suggests that the issue of multicollinearity has been improved.

Method

In this report, the Logistic regression model will be used for regression analysis, and a prediction classification model will be established. The Logistic regression model is mainly used to analyze the relationship between independent variables and discrete dependent variables.

Dependent variables are generally classified variables of "0-1" type. In this study, the main research content is personal credit risk evaluation, and the dependent variable y is binary variable with values of 0 and 1 respectively; y=1 represents a customer with default behavior, and y=0 represents a customer without default record. Logistic is essentially a discriminant model based on conditional probability. In actual credit approval classification, a threshold is set for classification, so Logistic regression model can also be regarded as a probability estimation, that is, an estimation of the default probability of the user.

ROC curve is used to judge the validity of the model, which represents the result combination of multiple confusion matrices. assuming that the threshold definition in the above model is unsuccessful, the prediction results of the model are simply sorted in descending order, and the threshold is defined by each probability value in sequence, so that many confusion matrices can be generated.

Confusion matrix is the basis of ROC curve drawing, and it is also the most basic, intuitive and simple method to measure the accuracy of classification model. Confusion matrix is to count the number of observed values of the wrong and right classes of the classification model respectively, and then display the results in a table.

Model fitting

First, we will look at the effectiveness of our continuous variable transformations. We will fit two models, one before transformations and one after transformations, and compare their respective AICs.

Table 2: AIC before and after transformations

Transformations	AIC
Before	3580.489
After	3439.412

As we can see from the table above, AIC for our model was lower after our proposed transformation. Hence, we will keep our transformation and go on to build our model. We will start from the full model (all variables are included in the model) and reduce variables until all coefficients of our variables are significant.

Table 3: Model variable coefficients

Variable	Coefficient
SEX	-1.197381
AGE	-0.170410
PAY_1	0.009844
BILL_AMT1	0.581024
PAY_AMT1	2.524044
PAY_AMT2	2.560198
PAY_AMT5	0.000003

After our step-by-step variable reduction, our final model has variables **SEX**, **AGE**, **BILL_AMT1**, **PAY_AMT1**, **PAY_AMT2** and **PAY_AMT5**. With only 6 variables, our model is also easy to interpret and have strong explanatory power. Comparing between the coefficients of variables **PAY_AMT**, we can see that recent paid amount are much important while the older ones are less important.

All of the coefficients of our variables are significant at the level of 0.1 as well. Using ANOVA to compare our initial model (full model) and our final model, we get a p-value of 0.001, which suggest that our changes are not significant and it appropriate that we removed our suggested variables. Furthermore, our selected variable model has a better goodness of fit, and AIC is smaller at 3439.2.

Results

Model diagnosis

Fitting the residual diagram of our model, the student residuals shows that there are some trends in the fitting residual of the model, but the overall fitting is good.

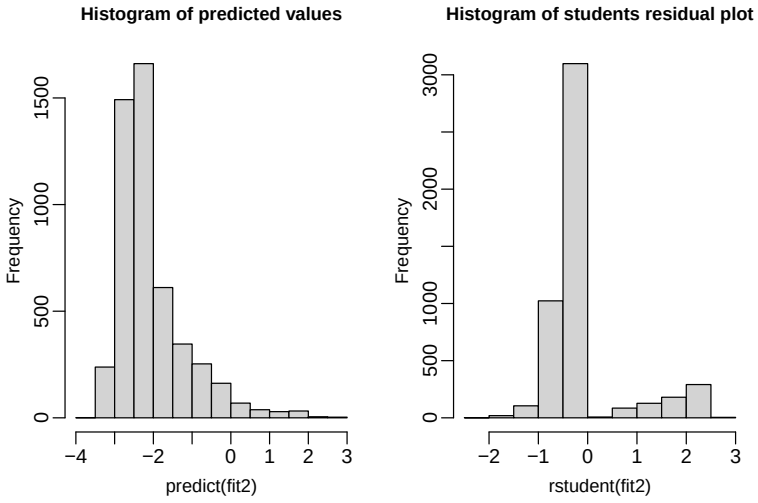

Figure 4: Histogram of Predicted value and Students residual

We looked into more residual plots such as deviance residual plot and pearson residual plot and they convey similar ideas of trend.

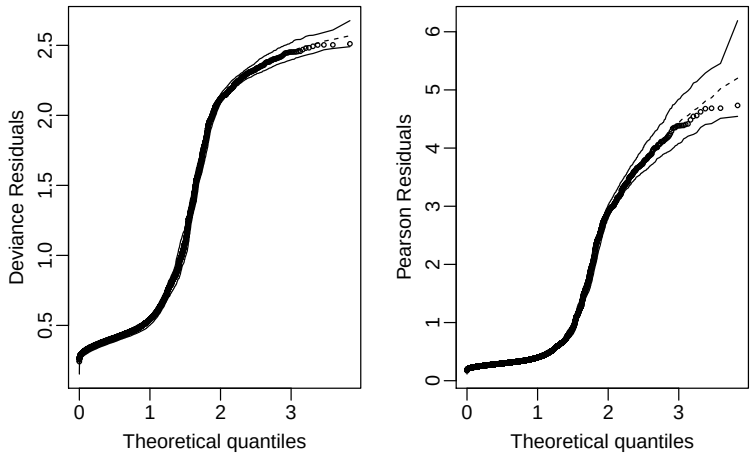

Figure 5: Deviance residual and Pearson residual plots

Looking into each variables, Added Variable Plots each variable is evenly distributed, and the residual of partial residual graph is evenly distributed on the trend line. All these suggest suitable fit of our model.

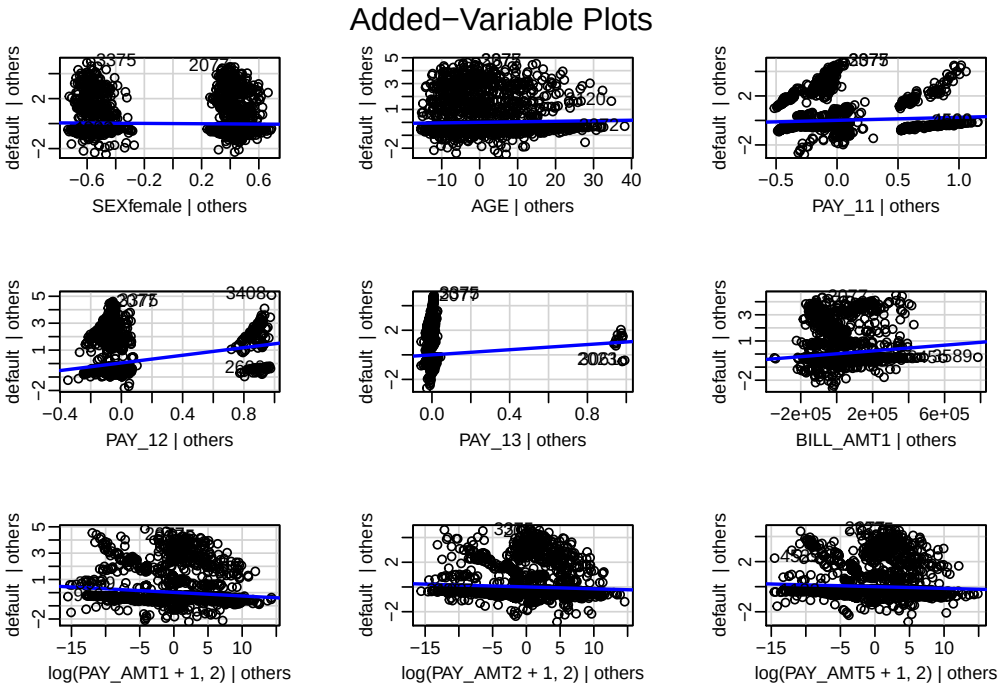

Figure 6: Added-Variables and Component-Residual Plots

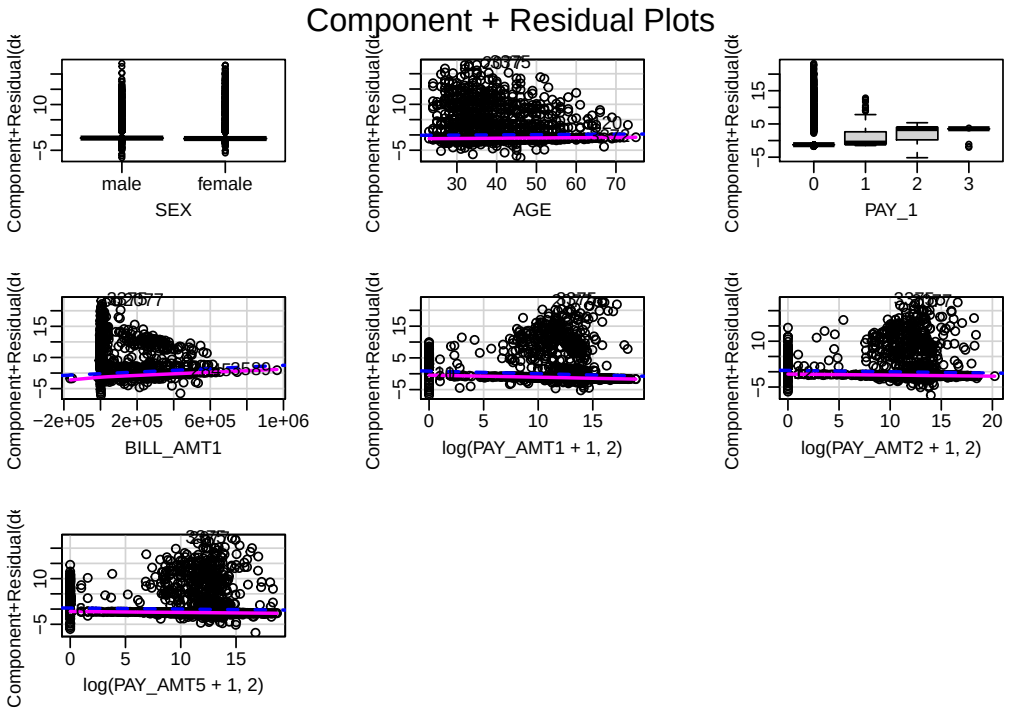

Figure 7: Added-Variables and Component-Residual Plots

Model prediction and classification

The purpose of establishing the model is to use the model to predict and classify credit card customers, and check whether customers have the possibility of default payment by fitting variables. The built model is used for logistics regression classification, and the confusion matrix is obtained by calculation. There are a large number of first-class error data and second-class error data. By selecting getting a loop to find optimal cutoff point for accuracy, we set our cutoff point is finally set at 0.55, so as to get the maximum accuracy. Then the accuracy, false positive rate and false negative rate are calculated, and the accuracy is 88.05%.

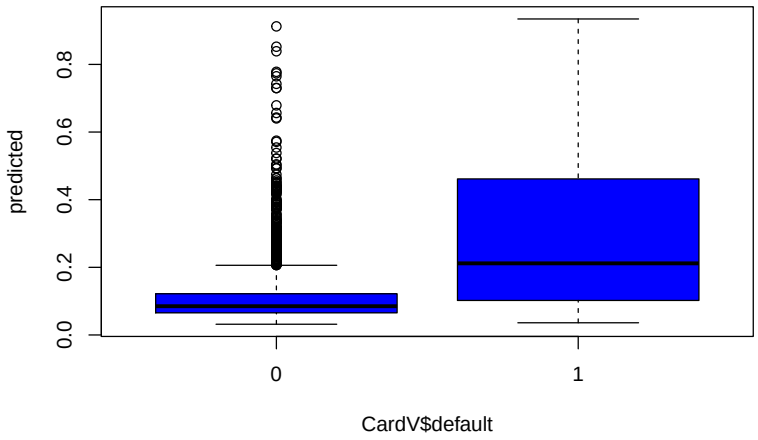

Figure 8: Boxplot to see probability and classfication

Below is our confusion matrix for when threshold = 0.55.

```
##       ypred
##        FALSE TRUE  Sum
##    0    1744   16 1760
##    1     229   62  291
##    Sum  1973   78 2051
```

However, we are not satisfied with this because our false negative is extremely high at 229 predictions, this means that our recall is low at merely 21.3%. High recall in our model is important as we want to be able to identify most or even all customers that will default, as these are customers that will create loss for the credit card company. Therefore, we will try find a suitable threshold that can manage the trade-off between accuracy and recall.

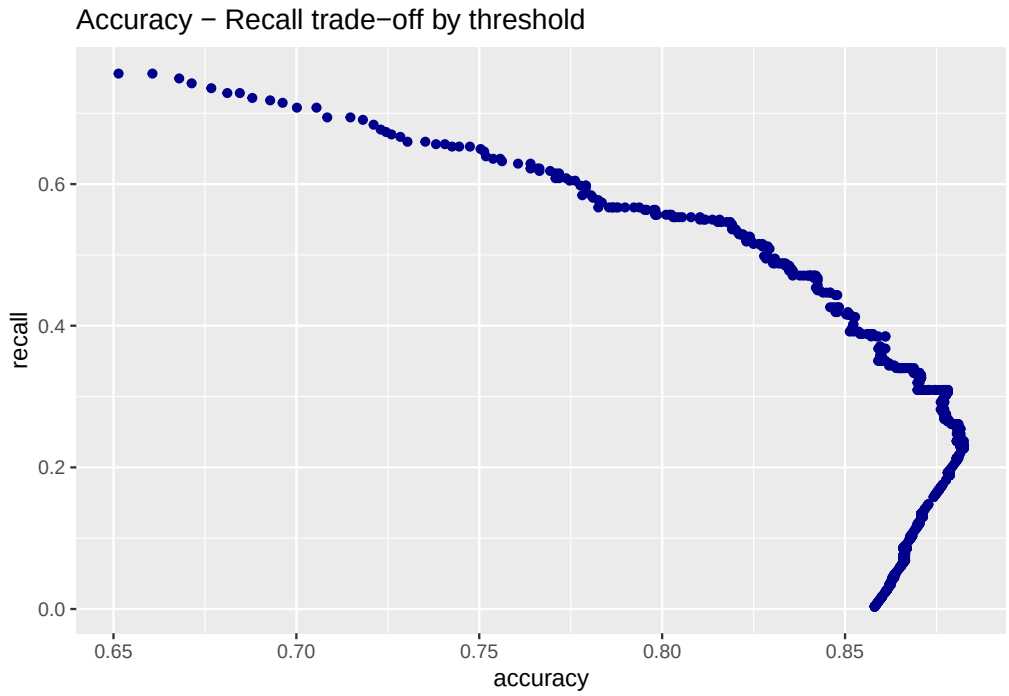

Figure 9: Accuracy and Recall trade-off plot

Using the graph above, we aim for a recall of above 0.6, which means that our model will be able to identify 60% of the default customers, we set our threshold as 0.146 and we get the following confusion matrix.

```
##       ypred
##        FALSE TRUE  Sum
##    0    1416  344 1760
##    1     115  176  291
##    Sum  1531  520 2051
```

With this threshold, our recall is 60.5% and our accuracy is 77.6%. We lost about 10% of our accuracy but we are now able to identify more than 60% of the defaulting customers instead of the initial 20 and I believe that it is a worthy trade-off.

In order to check the accuracy of classification, we fit the ROC curve. The closer the ROC curve is to the upper left corner, the higher the recall rate of the model. The point on the ROC curve closest to the upper left corner is the best threshold for the least classification errors, and the total number of false positive cases and false negative cases is the least.

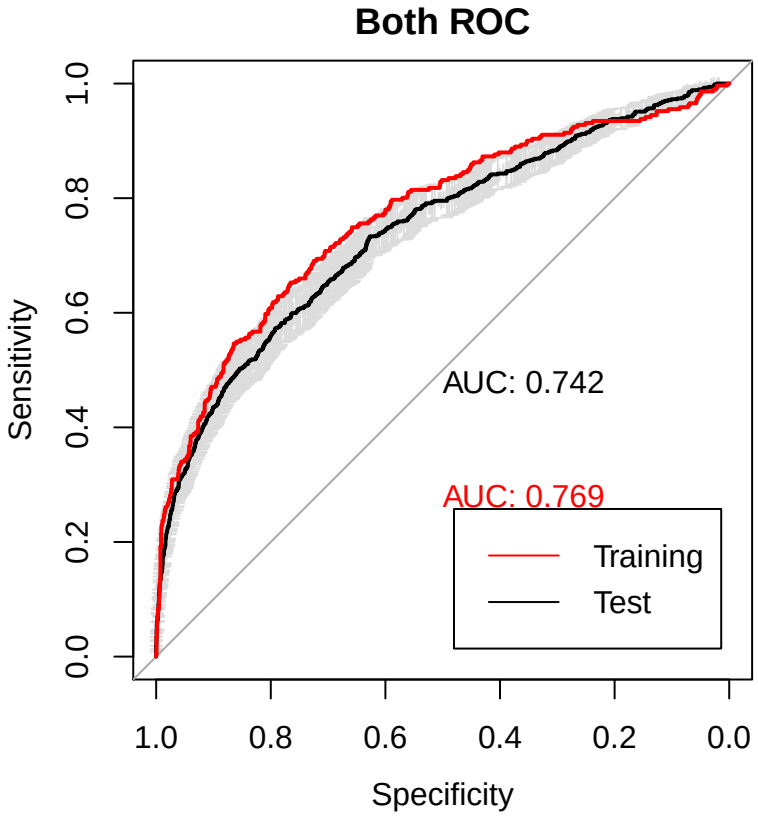

Figure 10: ROC plots on training and testing data set

The ROC for the training set is only for comparison purposes, it is also expected that our AUC for the training set is higher than that of the testing set and our model will be more fitted to our training data. Calculating our AUC, we get AUC = 0.742 with the 95% confidence interval being [0.7367,0.8013]. It can be concluded that AUC is average but definitely not ideal.

Limitations

The first limitations of our model is within the transformations applied to our predictor variables. While applying transformation, even though our predictor variables can become less skew and might lead to better model, the interpretability of our model is reduced in the process. This issue is further enhanced by the fact that we are using Logistic Regression model for prediction. Logistic Regression model, given by their nature of having discrete outcomes, they are less sensitive to skewed predictor variables, therefore, this further reduce the need for transformations.

In addition, in Logistic Regression, classification is done by manually selecting a threshold. While selected threshold might be good while being used on our training and testing data, it is not guarantee to be the optimal threshold for future data. This might reduce our model's effectiveness for future data. Instead, we can use more advance models for classification such as Decision Tree, Random Forest or even XGBoost. However, using more advance models can also lead to over-fitting but such issues will not be discussed in the scope of this paper.

Conclusion

In this paper, logistics regression is used to build the model, and the risk assessment of credit card applicants is carried out through data mining tools. In the process, the shortcomings of building the model are fully understood, and the model optimization is not good enough, and the prediction accuracy is not high enough. The data predicted by the test set shows that the classification results are not good enough, and the sampling method can be optimized, and stratified sampling and other methods can be considered. From the information of the variables in the article, it can be known that men, who have delayed repayment for more than one month, have a positive impact on default payment, and are more inclined to default payment, but these variables are not enough to build a complete model, so more effective variables can be considered to build a model. The credit card risk assessment method constructed in this paper cannot completely cover the credit assessment system, but is only a part of it. With the help of a more complete credit evaluation system, customer information can be obtained more effectively and comprehensively. With the popularization of credit card business and the continuous development of social credit, customer credit evaluation is becoming more and more complicated. At the same time, data mining tools are used more and more frequently, which is more helpful for us to complete the evaluation of customer credit. With the advantage of data mining, we can summarize the habits and laws of customer credit more efficiently, help banks to complete the mining of key elements, and help them to carry out customer credit information verification. At the same time, it also plays a guiding role in further perfecting the establishment of customer credit evaluation system.

Reference

[1]Altman E I .Financial Ratios,Discriminates Analysis and the Prediction of Corporate Bankruptcy[J].Journal of Finance,1968.

[2]Ausubei and Lawrence M.The Failure of Competition in the Credit Card Market[J].The American Economic Review,1991(81):50-81.

[3]Balse Committee on Banking Supervision.Credit risk modeling:current practices and applications,1999.

[4]Chen T,Guestrin C.XGBoost:A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM,2016:785-794.

[5]Jackson J R.Simluation Research on Job Shop Production[J].Naval Res Log Quart,1957,4(3):287-295.

[6]Mays E Handbook of Credit Scoring[M]. Fizroy Dearborn,2004.

Appendix

```
#Load libraries
library(ggplot2)
library(rpart)
library(car)
library(MASS)
library(Stat2Data)
library(plyr)
library(hnp)
library(DAAG)
library(sjPlot)
library(ROCR)
library(pROC)
library(VIM)
library(ggpubr)
library(kableExtra)
library(moments)
library(yardstick)

#Load data
CardT <- readRDS(file="CardT.rds")
CardV <- readRDS(file="CardV.rds")

#Check for missing values
allCards = rbind(CardT,CardV)
aggr(allCards, prop = FALSE, combined = TRUE, numbers = TRUE, srtVars = TRUE,
     sortCombs = TRUE, cex.axis = .7, oma = c(8,5,5,2))

#Analysis on missing values
attach(allCards)
allCards$EducationNA <- ifelse(is.na(EDUCATION), "Missing", "Not missing")
allCards$MarriageNA <- ifelse(is.na(MARRIAGE), "Missing", "Not missing")

attach(allCards)
par1<- par(mfrow=c(2,2), mar =c(3,3,3,3))
boxplot(LIMIT_BAL~EducationNA, col=c("red", "light blue"), xlab = "", ylab="")
mtext(text = "Education", side =1, line =2, cex=.7)
mtext(text = "Credit Limit", side =2, line =2, cex = .7)
boxplot(AGE~EducationNA, col=c("red", "light blue"), xlab = "", ylab="")
mtext(text = "Education", side =1, line =2, cex=0.7)
mtext(text = "Age", side =2, line =2, cex = 0.7)
mosaicplot(table(EducationNA, SEX), cex.axis = 0.55, col = c("red", "light blue"),
           main = "", las=1)
par(par1)

#Omit missing values
```

```
CardT <- na.omit(CardT)
CardV <- na.omit(CardV)

#Transform our factor variables into factor
CardT$SEX<-as.factor(CardT$SEX)
CardT$EDUCATION<-as.factor(CardT$EDUCATION)
CardT$MARRIAGE<-as.factor(CardT$MARRIAGE)
CardT$default<-as.factor(CardT$default)
CardT$PAY_1<-as.factor(CardT$PAY_1)
CardT$PAY_2<-as.factor(CardT$PAY_2)
CardT$PAY_3<-as.factor(CardT$PAY_3)
CardT$PAY_4<-as.factor(CardT$PAY_4)
CardT$PAY_5<-as.factor(CardT$PAY_5)
CardT$PAY_6<-as.factor(CardT$PAY_6)

#Density plots to check for skewness
LimitBalPlot <- ggplot(data = CardT,aes(log(LIMIT_BAL,2),fill=factor(default))) +
  geom_density(alpha=.6)
AgePlot <- ggplot(data = CardT,aes(AGE,fill=factor(default))) +
  geom_density(alpha=.6)
BillAmtPlot <- ggplot(data = CardT,aes(BILL_AMT1,fill=factor(default))) +
  geom_density(alpha=.6)
PayAmtPlot <- ggplot(data = CardT,aes(log(PAY_AMT1+1,2),fill=factor(default))) +
  geom_density(alpha=.6)

ggarrange(LimitBalPlot,AgePlot,BillAmtPlot,PayAmtPlot,ncol = 2,nrow = 2)

#Factor transformations
#For training data
CardT$PAY_1[CardT$PAY_1=="-1"]<-0
CardT$PAY_1[CardT$PAY_1=="-2"]<-0
CardT$PAY_2[CardT$PAY_2=="-1"]<-0
CardT$PAY_2[CardT$PAY_2=="-2"]<-0
CardT$PAY_3[CardT$PAY_3=="-1"]<-0
CardT$PAY_3[CardT$PAY_3=="-2"]<-0
CardT$PAY_4[CardT$PAY_4=="-1"]<-0
CardT$PAY_4[CardT$PAY_4=="-2"]<-0
CardT$PAY_5[CardT$PAY_5=="-1"]<-0
CardT$PAY_5[CardT$PAY_5=="-2"]<-0
CardT$PAY_6[CardT$PAY_6=="-1"]<-0
CardT$PAY_6[CardT$PAY_6=="-2"]<-0
CardT$PAY_1[CardT$PAY_1=="4"]<-3
CardT$PAY_1[CardT$PAY_1=="5"]<-3
CardT$PAY_1[CardT$PAY_1=="6"]<-3
CardT$PAY_1[CardT$PAY_1=="7"]<-3
CardT$PAY_1[CardT$PAY_1=="8"]<-3
```

```r
#For testing data
CardV$PAY_1[CardV$PAY_1=="-1"]<-0
CardV$PAY_1[CardV$PAY_1=="-2"]<-0
CardV$PAY_2[CardV$PAY_2=="-1"]<-0
CardV$PAY_2[CardV$PAY_2=="-2"]<-0
CardV$PAY_3[CardV$PAY_3=="-1"]<-0
CardV$PAY_3[CardV$PAY_3=="-2"]<-0
CardV$PAY_4[CardV$PAY_4=="-1"]<-0
CardV$PAY_4[CardV$PAY_4=="-2"]<-0
CardV$PAY_5[CardV$PAY_5=="-1"]<-0
CardV$PAY_5[CardV$PAY_5=="-2"]<-0
CardV$PAY_6[CardV$PAY_6=="-1"]<-0
CardV$PAY_6[CardV$PAY_6=="-2"]<-0
CardV$PAY_1[CardV$PAY_1=="4"]<-3
CardV$PAY_1[CardV$PAY_1=="5"]<-3
CardV$PAY_1[CardV$PAY_1=="6"]<-3
CardV$PAY_1[CardV$PAY_1=="7"]<-3
CardV$PAY_1[CardV$PAY_1=="8"]<-3

#Using VIF to check for correlation
model1 <- lm(default ~ log(LIMIT_BAL,2) + AGE + BILL_AMT1 + BILL_AMT2 +
               BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + BILL_AMT6 +
               log(PAY_AMT1+1,2) + log(PAY_AMT2+1,2) + log(PAY_AMT3+1,2) +
               log(PAY_AMT4+1,2) + log(PAY_AMT5+1,2) + log(PAY_AMT6+1,2),
             data = CardT)

Variable<- c("LIMIT_BAL","AGE","BILL_AMT1","BILL_AMT2","BILL_AMT3","BILL_AMT4",
             "BILL_AMT5","BILL_AMT6","PAY_AMT1","PAY_AMT2","PAY_AMT3","PAY_AMT4",
             "PAY_AMT5","PAY_AMT6")

Coefficient = c()
for (i in 1:14) {
  Coefficient[i] <- round(vif(model1)[i], digits=4)
}
df <- data.frame(Variable, Coefficient)

model1 <- lm(default ~ log(LIMIT_BAL,2) + AGE + BILL_AMT1 + BILL_AMT6 +
               log(PAY_AMT1+1,2) + log(PAY_AMT2+1,2) + log(PAY_AMT3+1,2) +
               log(PAY_AMT4+1,2) + log(PAY_AMT5+1,2) + log(PAY_AMT6+1,2),
             data = CardT)

Variable<-c("LIMIT_BAL","AGE","BILL_AMT1","BILL_AMT6","PAY_AMT1","PAY_AMT2",
            "PAY_AMT3","PAY_AMT4","PAY_AMT5","PAY_AMT6")

Coefficient = c()
for (i in 1:10) {
  Coefficient[i]<-round(vif(model1)[i], digits=4)
```

```
}
df2<-data.frame(Variable, Coefficient)
kable(list(df,df2), caption = "VIF before and after variable selection") %>%
  kable_styling(position = "center") %>%
  kable_styling(latex_options = "HOLD_position")

#Removing correlated variables
CardT <- subset(CardT,select=-c(BILL_AMT2,BILL_AMT3,BILL_AMT4,BILL_AMT5))

#Fitting models before and after transformations
fit0<-glm(formula = default ~ LIMIT_BAL + SEX + AGE + MARRIAGE + PAY_1 + PAY_2 +
            PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT6 + PAY_AMT1 +
            PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6,
          family = binomial(), data=CardT)

fit1<-glm(formula = default ~ log(LIMIT_BAL,2) + SEX + AGE + MARRIAGE + PAY_1 +
            PAY_2 +PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT6 +
            log(PAY_AMT1+1,2) + log(PAY_AMT2+1,2) + log(PAY_AMT3+1,2) +
            log(PAY_AMT4+1,2) + log(PAY_AMT5+1,2) + log(PAY_AMT6+1,2),
          family = binomial(), data=CardT)

df0 <- data.frame(c("Before","After"), c(AIC(fit0),AIC(fit1)))
colnames(df0) <- c("Transformations","AIC")
kable(df0, caption = "AIC before and after transformations") %>%
  kable_styling(position = "center") %>%
  kable_styling(latex_options = "HOLD_position")

#Fitting our final model
fit2<-glm(formula = default ~ SEX + AGE + PAY_1 + BILL_AMT1 + log(PAY_AMT1+1,2) +
            log(PAY_AMT2+1,2) + log(PAY_AMT5+1,2),
          family = binomial(), data=CardT)
summary(fit2)
Anova(fit2)

anova(fit1,fit2, test="Chisq")

#Model results, coefficient of variables
Variable<-c("SEX","AGE","PAY_1","BILL_AMT1","PAY_AMT1","PAY_AMT2","PAY_AMT5")
Coefficient = c()
for (i in 1:7){
  if (coef(fit2)[i]==0) {
    Coefficient[i] = "."
  } else {
  Coefficient[i] = round(coef(fit2)[i], digits=6)
  }
}
df = data.frame(Variable, Coefficient)
```

```r
kable(df, caption = "Model variable coefficients") %>%
  kable_styling(position = "center") %>%
  kable_styling(latex_options = "HOLD_position")

#Residual plots for model diagnostics
par0 <- par(mfrow=c(1,2),mgp=c(1.7,0.5,0),mar=c(3.5,3.5,3,0.5))
pred <- hist(predict(fit2), main = "Histogram of predicted values", cex.main=0.9,
             cex.lab=0.9)
students <- hist(rstudent(fit2), main = "Histogram of students residual plot",
                 cex.main=0.9,cex.lab=0.9)
par(par0)

par1 <- par(mfrow=c(1,2),mgp=c(1.7,0.5,0),mar=c(3.5,3.5,3,0.5))
hnp(fit2,resid.type="deviance",ylab="Deviance Residuals")
hnp(fit2,resid.type="pearson", ylab="Pearson Residuals")
par(par1)

par2 <- par(mfrow=c(1,2),mgp=c(1.7,0.5,0),mar=c(3.5,3.5,3,0.5))
avPlots(fit2) #Added Varaible Plots examples
crPlots(fit2,id=TRUE) #Partial Residual Plots
par(par2)

#Model Performance
predicted <- predict(fit2, CardV, type='response')
boxplot(predicted~CardV$default, col="blue")

#Confusion matrix for threshold = 0.55
ypred <- predicted > 0.55
addmargins(table(CardV$default, ypred))

#Searching for better thresholds
c = seq(from=0.1, to=0.9, by=0.001)
accuracy = c()
recall = c()
for (i in 1:length(c)){
  ypred <- predicted > c[i]
  confusionmatrix <- addmargins(table(CardV$default, ypred))
  accuracy[i] = (confusionmatrix[1,1]+confusionmatrix[2,2])/confusionmatrix[3,3]
  recall[i] = (confusionmatrix[2,2]/(confusionmatrix[2,1]+confusionmatrix[2,2]))
}

df = data.frame(accuracy,recall)
ggplot(data=df, aes(x=accuracy,y=recall)) +
  geom_point(color='darkblue') +
  labs(title="Accuracy - Recall trade-off by threshold")
```

```r
#Confusion matrix for threshold = 0.146
ypred <- predicted > 0.146
addmargins(table(CardV$default, ypred))

#ROC Curve
par(pty="s")
plot.roc(CardT$default,predict(fit2, CardT, type="response"),ci=TRUE,
         of="thresholds",ci.type="shape",print.auc=TRUE, main="Both ROC")
plot.roc(CardV$default,predict(fit2, CardV, type='response'),add=TRUE,col="red",
         print.auc=TRUE, print.auc.x=0.5, print.auc.y=0.3,print.auc.col="red")
legend("bottomright",legend=c("Training","Test"),col=c("red","black"),lty=c(1,1),
       inset=c(0.05,0.05))
```