# BT2103 Project 3

Chen Haoli, Lo Zhi Hao, Luah Jun Yang, Toh Zhan Ting

## Table of Contents

## Introduction to the Dataset and Problem Statement

The cash and credit card debt problem that Taiwan's credit card issuers experienced in recent years is predicted to peak in the third quarter of 2006 (Chou, 2006). Taiwan's card-issuing banks over-issued cash and credit cards to unqualified applicants in an effort to gain market dominance. In addition, most cardholders, regardless of their capacity to pay back, abused their cards for consumption and racked up large credit and cash-card debt. The crisis damaged consumer confidence in finance, and therefore presents a significant problem for both banks and cardholders.

Crisis management and risk prediction take place upstream and downstream in a mature financial system, respectively. The main goal of risk prediction is to lessen the harm and uncertainty caused by corporate performance or individual customer credit risk by using financial information, such as business financial statements, customer transaction and repayment histories, etc.

Therefore, with extensive data collected from the period of April to September in 2005, this report aims to build a predictive model to accurately forecast clients who have tendencies to default on the bank and thus ensures the profitability and stability of banks. In Finance, a default occurs when a borrower doesn't fulfill the terms of the loan. In this situation, default would occur if the cardholder failed to pay the credit card account within a given month.

In the dataset, there are 23 explanatory variables, together with a dependent variable of client's default status. The detailed description is as follows:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

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.

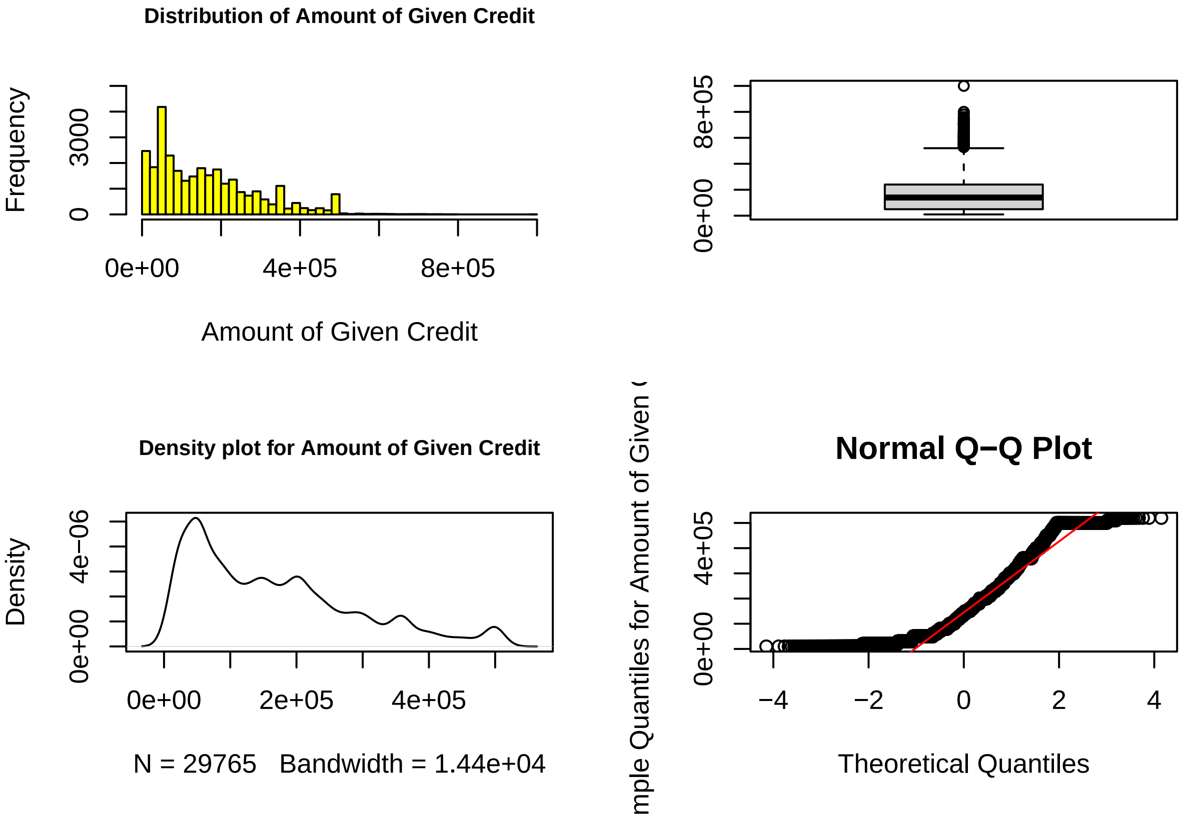
X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005;. . .;X23 = amount paid in April, 2005.

# Exploratory Data Analysis

As all of the columns in the data are originally categorized as characters, in order to proceed with the data pre-processing and data visualization, we decided to transform the columns that are considered as continuous and numeric back to a numeric data type. After further research from reading through the description of our data, we identified x1, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16, x17, x18, x19, x20, x21, x22, x23 as the data columns that we may need to transform. x2, x3, x4 are factors. For variables x3 and x4, we dropped the factor levels 0, which are indicative of N.A values for those variables Education and Marital Status respectively and they only amount to a small number of outliers. The other non-classified factor levels for these 2 variables are parked under the 'others' factor so that we will not exclude too many data points.

By using summary statistics and graphical representations, we are conducting preliminary analyses on the data in order to find trends, identify anomalies, inconsistencies, and missing values to test hypotheses and verify assumptions.

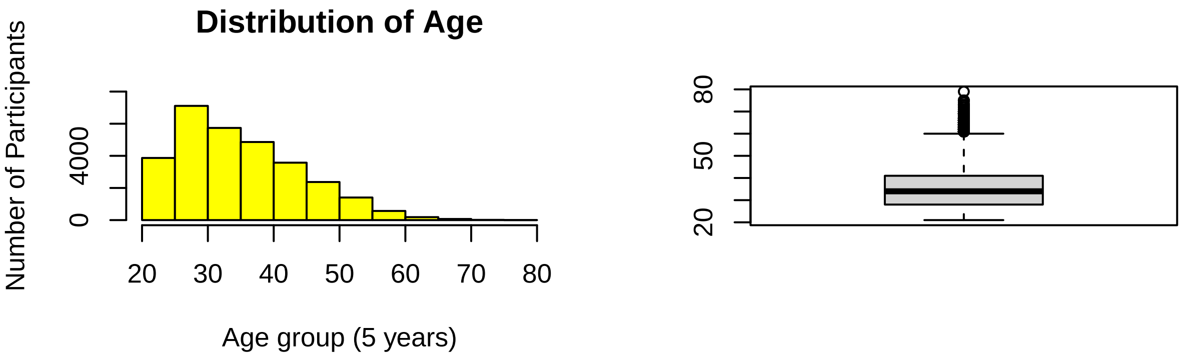
## Amount of Given Credit (X1)



From the boxplot, we observed points lie outside Q3 + 1.5IQR from the mean. We consider these points as outliers and we will remove these points from dataset From the density plot, we can observe that the data is positively skewed and it is not normally distributed.
From the qqplot, we can observe that the data is not normally distributed since the line and plots are not truly aligned.

Since the amount of given credit is positively skewed, we will use log transformation to improve the distribution of the data to normality.

## Age (X5)



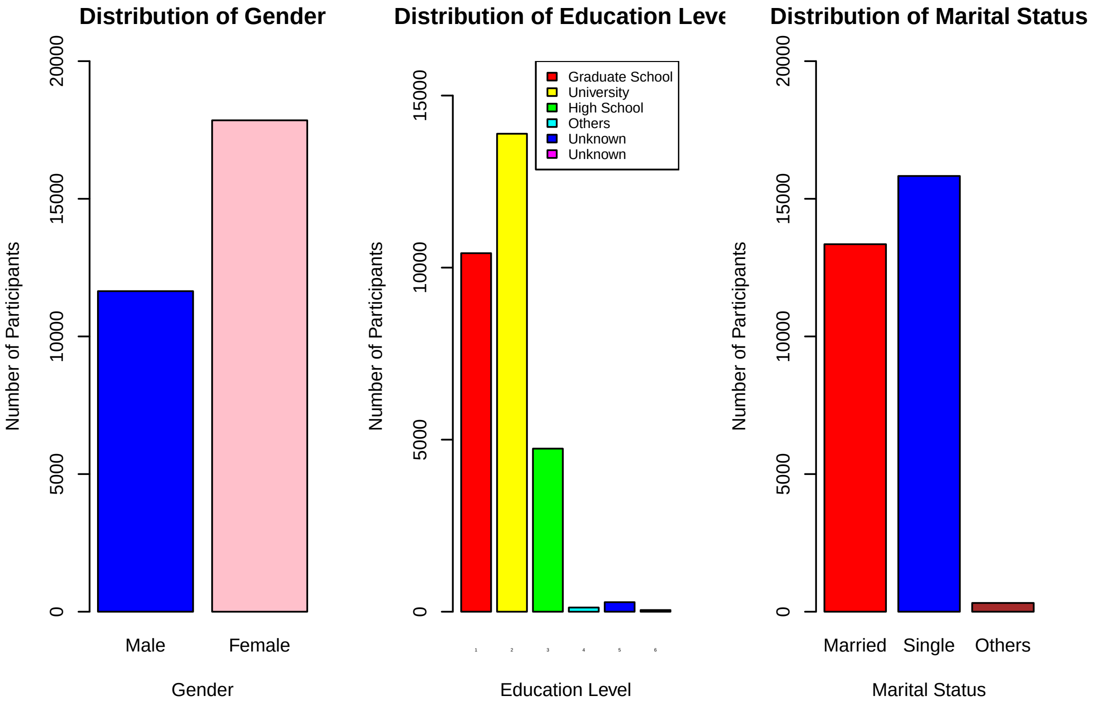
From the boxplot, we observed points lie outside Q3 + 1.5IQR from the mean. We consider these points as outliers and we will remove these points from dataset

## Gender (X2), Education (X3), and Marital Status (X5)

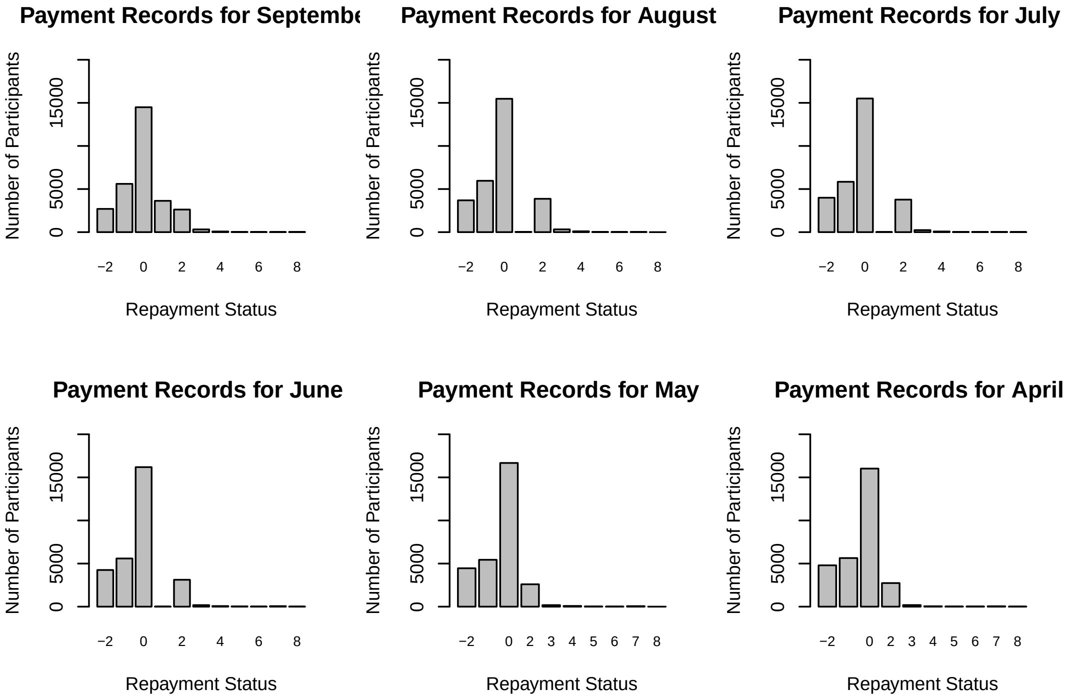
```
##   Male Female
##  11645  17851
```

```
## Graduate School       University     High School           Others           Unknown
##           10419            13890            4740              122               277
##         Unknown
##              48
```

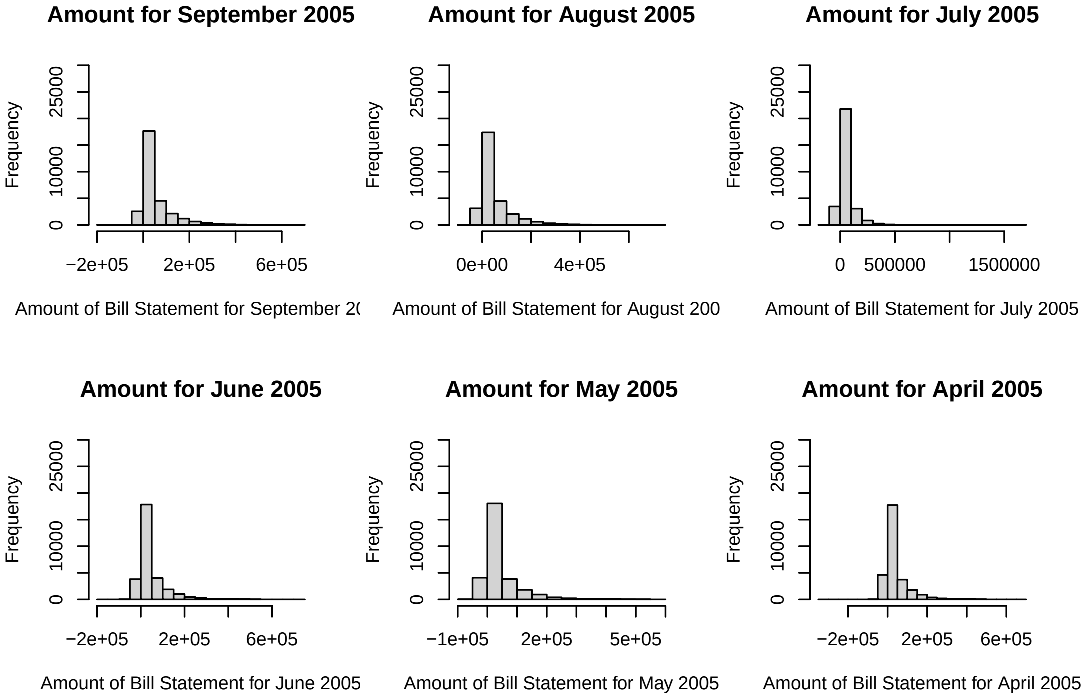
```
## Married  Single  Others
##   13350   15828     318
```



## Monthly Payment Records for 2005 (X6 - X11)

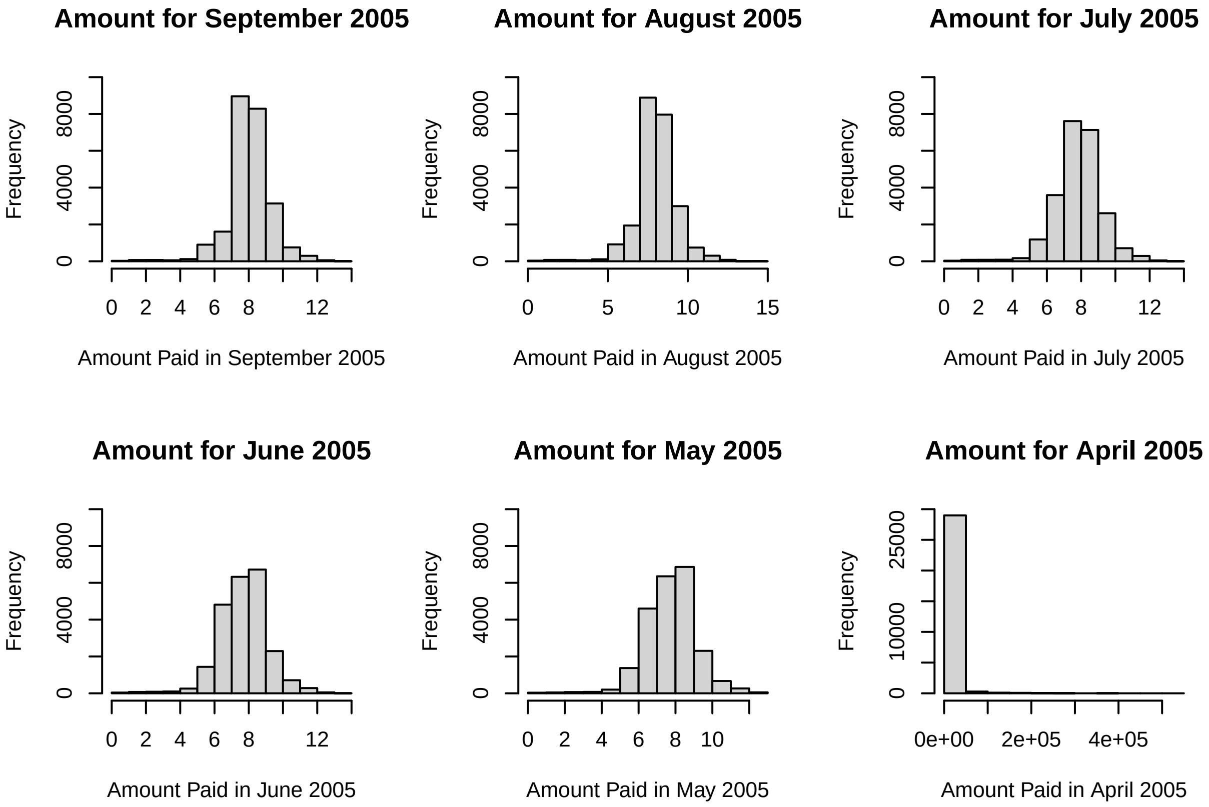


## Monthly Bill Distributions for 2005 (X12 - X17)



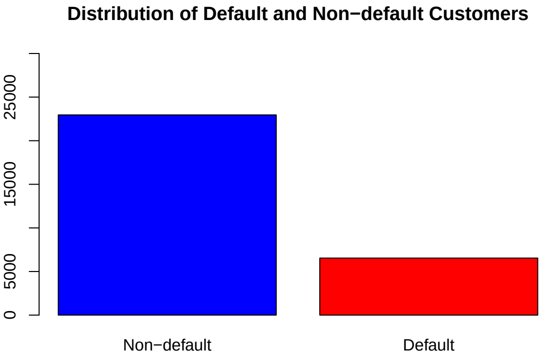
## Amount of previous payment (X18 - X23)

We will perform log transformation for Amount Paid in order to make it resemble a normal distribution.



## Distribution of Default and Non-default data

```
## Non-default     Default
##       22957        6539
```



### Data Pre-Processing

As we noticed from the data visualisation shown in our part above, there is a significant issue with the data we gathered from our dataset, as some of the data are not described in our original cited data source.

For example, for distribution of education level and marital status, we noticed some entries that does not belong to any factor levels that were described in the original data source. Similarly, for the history of payment records, we also realized the occurrence of erroneous factor levels such as -2 and 0, which are not described for in the original data source.

First, we are going to change the names of each columns for better understanding and easier interpretation in our following chapters.

Moving on we are going to perform some data manipulation on some of the data anomalies that we spotted earlier. In that sense, we are going to manipulate and change the data points that are originally not described in our cited data source.

### Data Manipulation for Marital Status (X2)

As we can see from the visualisation for distribution of marital status for our clients, there are categories that are not described for in our original data cited source, namely 0. For our team, as we consider 0 values as possible null values where our clients did not provide their marital status information properly, we decided to drop rows with 0 values recorded for marital status.

```
##
##     1     2     3
## 13350 15828   318
```

### Data Manipulation for Education Status (X3)

As we can see from the visualisation for distribution of education level for our clients, there are categories that are not described for in our original data cited source, namely 0, 5 and 6. For our team, as we consider 0 as a possible null value where our clients did not provide their academic credentials, we decided to drop those observations with 0 recorded as their education level. Meanwhile, as 5 and 6 may be education levels that are higher than or not included in the provided options (such as PhD), we decided to manipulate those data to add them into the 'others' category, 4 that is provided.

```
##
##     1     2     3     4
## 10419 13890  4740   447
```

## Feature Selection

Make use of feature selection methodologies to select the most relevant independent variables to create a prediction model.

For our project, as it is a large dataset with around 30,000 observations, for each type of feature selection method, we are going to perform a test in order to determine the best features in that category, namely:

Forward and Backward variable selection

Filter Method - ANOVA for continuous data and Chi Squared Test for categorical data

Wrapper Method - Boruta Method

Embedded Method - Information Gain

**Forward Variable Selection:**

- Start with model containing no possible explanatory variable and for each variable in turn, we will investigate effect of adding variable from the current model.
- 5 most significant variables will be used for our subsequent models

```
##   (Intercept) Amt_credit Gender Edu_Lvl Marital Age PayRec_Sep PayRec_Aug
## 1           1          0      0       0       0   0          1          0
## 2           1          1      0       0       0   0          1          0
## 3           1          1      0       0       0   0          1          0
## 4           1          1      0       0       0   0          1          1
## 5           1          1      0       0       1   0          1          1
##   PayRec_Jul PayRec_Jun PayRec_May PayRec_Apr BillAmt_Sep BillAmt_Aug
## 1          0          0          0          0           0           0
## 2          0          0          0          0           0           0
## 3          0          0          0          0           1           0
## 4          0          0          0          0           1           0
## 5          0          0          0          0           1           0
##   BillAmt_Jul BillAmt_Jun BillAmt_May BillAmt_Apr PaidAmt_Sep PaidAmt_Aug
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
##   PaidAmt_Jul PaidAmt_Jun PaidAmt_May PaidAmt_Apr   rsq adjr2      cp       bic
## 1           0           0           0           0 0.106 0.106 659.541 -3290.498
## 2           0           0           0           0 0.113 0.113 422.683 -3514.801
## 3           0           0           0           0 0.117 0.116 315.037 -3612.826
## 4           0           0           0           0 0.121 0.120 179.525 -3738.945
## 5           0           0           0           0 0.122 0.122 129.422 -3780.503
##        rss
## 1 4548.953
## 2 4512.917
## 3 4496.375
## 4 4475.628
## 5 4467.768
```

**Backward Variable Selection:**

- Start with model containing all possible explanatory variables and for each variable in turn, we will investigate effect of removing that variable from the current model.
- 5 most significant variables will be used for our subsequent models

```
##   (Intercept) Amt_credit Gender Edu_Lvl Marital Age PayRec_Sep PayRec_Aug
## 1           1          0      0       0       0   0          1          0
## 2           1          0      0       0       0   0          1          0
## 3           1          0      0       0       0   0          1          1
## 4           1          1      0       0       0   0          1          1
## 5           1          1      0       0       0   1          1          1
##   PayRec_Jul PayRec_Jun PayRec_May PayRec_Apr BillAmt_Sep BillAmt_Aug
## 1          0          0          0          0           0           0
## 2          0          0          0          0           1           0
## 3          0          0          0          0           1           0
## 4          0          0          0          0           1           0
## 5          0          0          0          0           1           0
##   BillAmt_Jul BillAmt_Jun BillAmt_May BillAmt_Apr PaidAmt_Sep PaidAmt_Aug
## 1           0           0           0           0           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
##   PaidAmt_Jul PaidAmt_Jun PaidAmt_May PaidAmt_Apr   rsq adjr2      cp       bic
## 1           0           0           0           0 0.106 0.106 659.541 -3290.498
## 2           0           0           0           0 0.113 0.113 426.521 -3511.017
## 3           0           0           0           0 0.119 0.119 230.592 -3696.520
## 4           0           0           0           0 0.121 0.120 179.525 -3738.945
## 5           0           0           0           0 0.122 0.122 130.981 -3778.950
##        rss
## 1 4548.953
## 2 4513.496
## 3 4483.635
## 4 4475.628
## 5 4468.003
```

From both the Forward and Backward Variable Selection, we observed similarities in the significant variables selected in both models (i.e.,Amt_credit, PayRec_Sep, PayRec_Aug, PayRec_Jul). However, we also observed that Marital Status & Age are ranked differently in both models. Thus, in order to increase accuracy of our subsequent models, we decided to include both Marital & Age.

## Splitting into train and test set

First, we will split our dataset into train and test set before actually performing the feature selection models on our data. This is to prevent any leakage of information from our test set into our training set leading to biased and overfitted models.

## Anova Test

```
##       Variable      p-value
## 2          Age 0.000000e+00
## 14 PaidAmt_Apr 0.000000e+00
## 8  BillAmt_Apr 5.897689e-20
## 10 PaidAmt_Aug 4.891178e-15
## 11 PaidAmt_Jul 8.157781e-13
## 12 PaidAmt_Jun 1.371457e-12
## 9  PaidAmt_Sep 1.665843e-11
## 1   Amt_Credit 2.898041e-11
## 13 PaidAmt_May 2.794714e-10
## 3  BillAmt_Sep 6.482189e-03
## 4  BillAmt_Aug 1.310158e-02
## 5  BillAmt_Jul 7.495121e-02
## 6  BillAmt_Jun 1.744674e-01
## 7  BillAmt_May 1.864872e-01
```

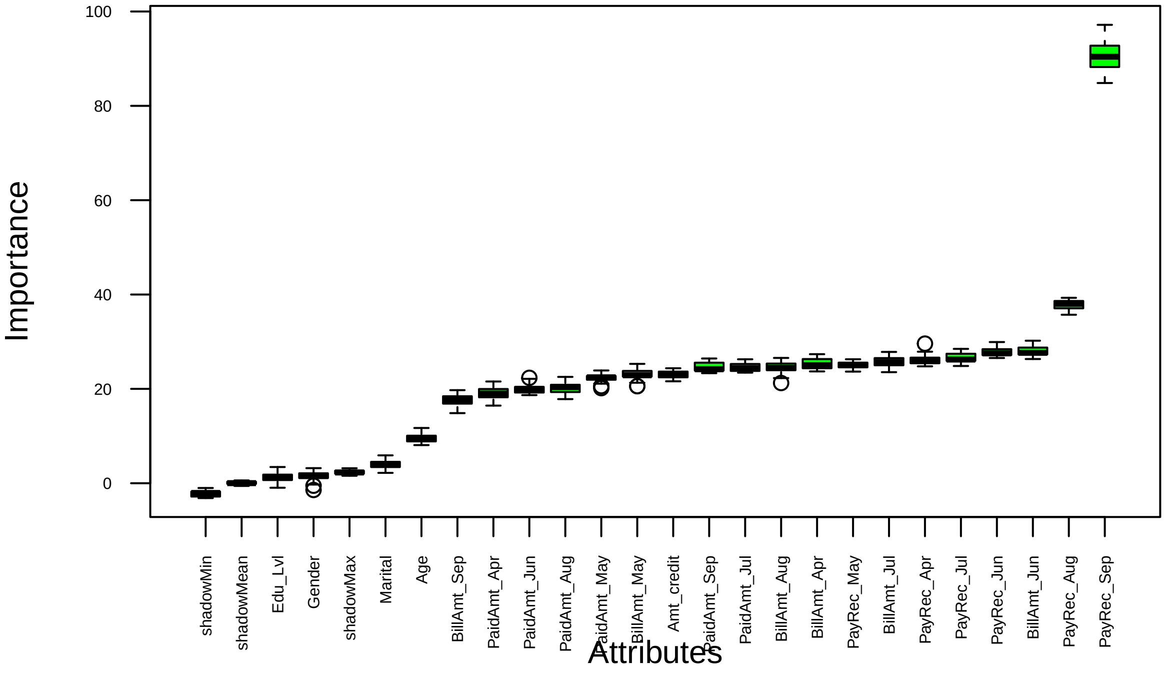
## Chi Square test

Next, for categorical variables, Chi Squared Method is used to identify the most important features.

```
##   Statistics      p-value            V3
## 1     Gender     6.713353  3.485089e-02
## 2    Edu_Lvl    85.508305  2.492442e-05
## 7 PayRec_Jun   983.964863 4.943317e-206
## 6 PayRec_Jul  1054.926738 2.457447e-221
## 5 PayRec_Aug  1131.704234 7.713010e-237
## 4 PayRec_Sep  1303.027160 8.499636e-274
## 3    Marital  1812.283792  0.000000e+00
## 8 PayRec_May 11924.409892  8.638387e-01
```

### Wrapper method for feature selection

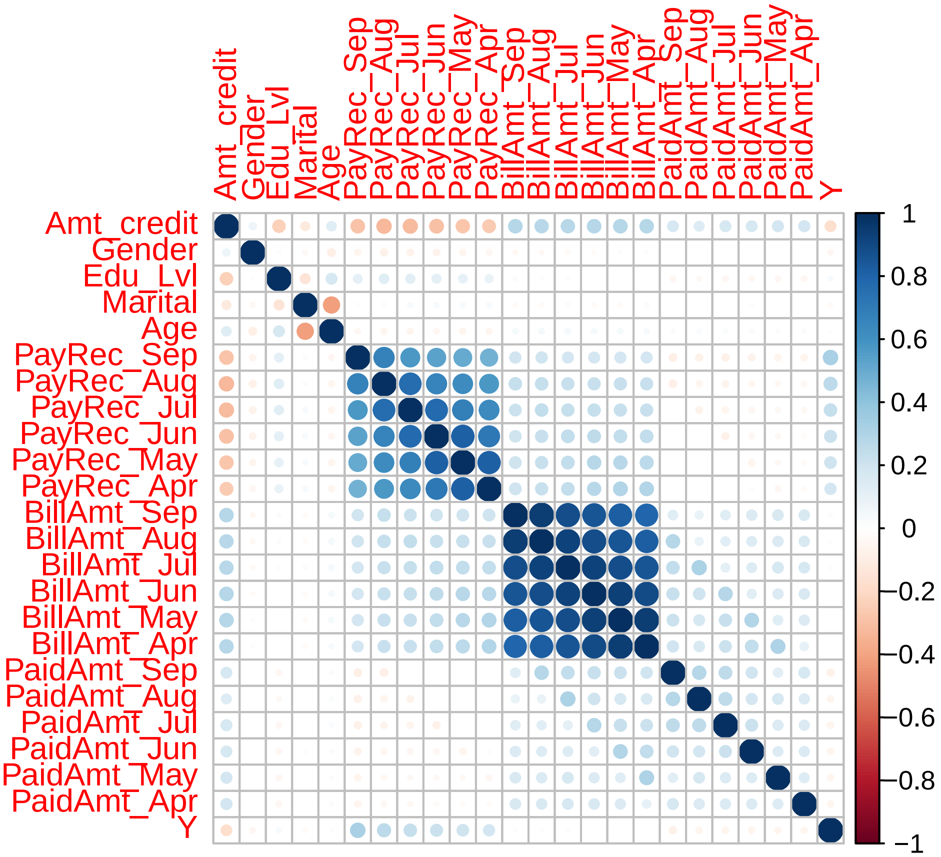
For wrapper method, the package and method that we are going to use is the Boruta Method that utilises random forest decision tree model in computing for the importance of each feature.



### Embedded method for feature selection

We utilise the information gain function from the FSelectorRcpp package to inspect and identify the important features in our model.

### Correlation

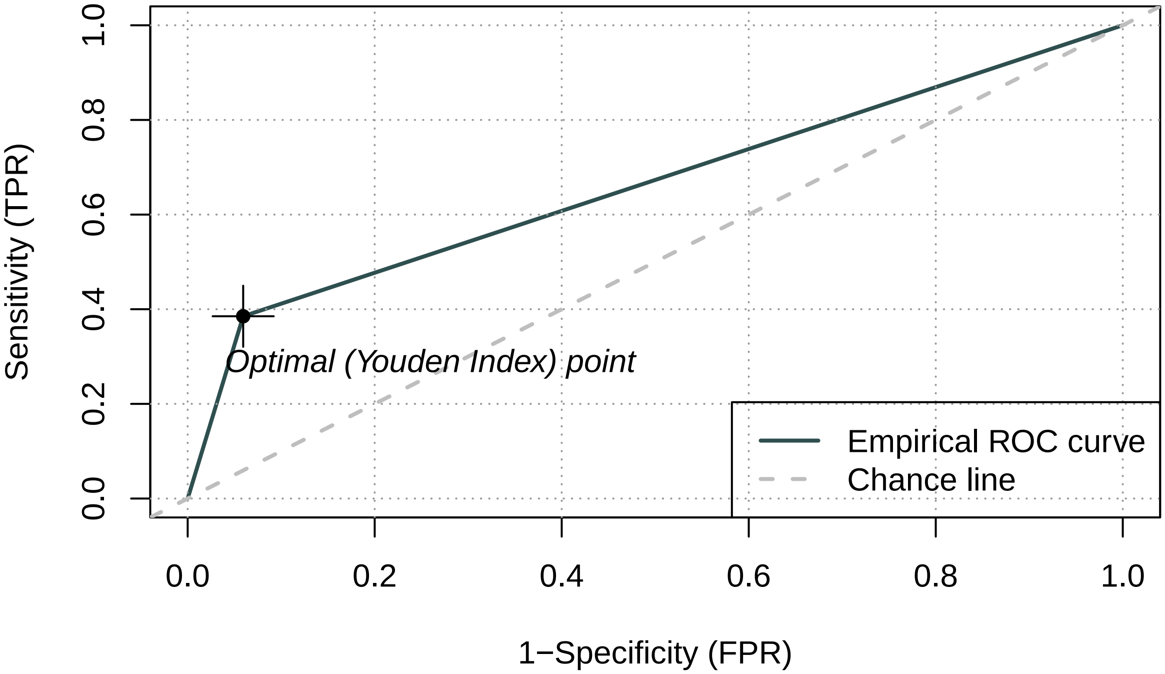


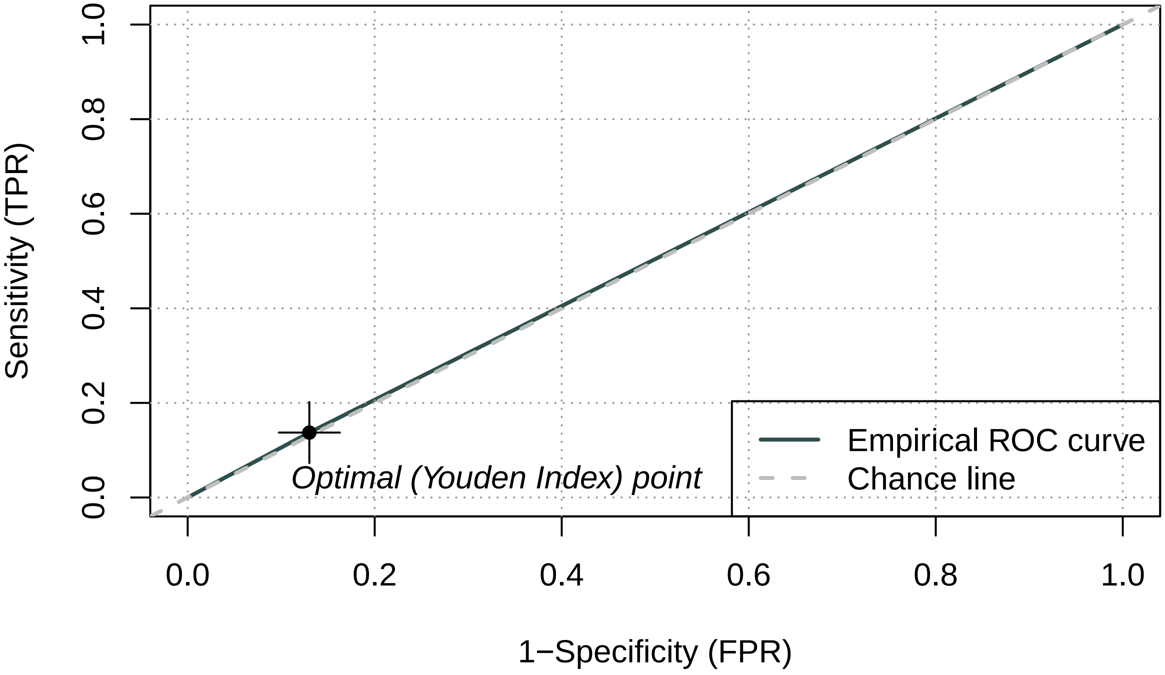
The highly correlated attribtues are BillAmt_Jun, BillAmt_May, BillAmt_Jul, BillAmt_Apr, BillAmt_Aug, PayRec_Jun, PayRec_Jul, PayRec_Aug, PayRec_May.

### Model Selection

For our models, a way to identify the effectiveness and accuracy of each of our models is to evaluate each of the models based on accuracy, null accuracy, ROC and AUC values and harmonic mean. Each of these methods have their own merits and advantages, and by computing the confusion matrix and compute each of these values, we can get a deeper understanding on how each of our models are performing.
For null accuracy, which is the accuracy that could be achieved by always predicting the most frequent class, is used as a metrics for reference for the effectiveness of the model on overall. If a model performs better than the null accuracy significantly, it suggests that the model itself is effective and should be utilises in our decision making.
For harmonic mean, as it is a function of both recall and precision, it therefore strives and performs better in imbalanced datasets, where the accuracy score may be affected by the large number of data samples on one side. Therefore, we decided to include harmonic mean as a way to evaluate our model performance.
For ROC/AUC curves, the curve measures the sensitivity and specificity of the model, and provides us with a clearer view and understanding on our final model. It also serves as a way to prevent overfitting from happening in our model. Essentially, this visualisation method provides us with a clearer understanding on the performance of our models, and thus is included as a way to evaluate our model.
From our calculations, our null accuracy for the dataset is at 0.778.

### Logistic Regression





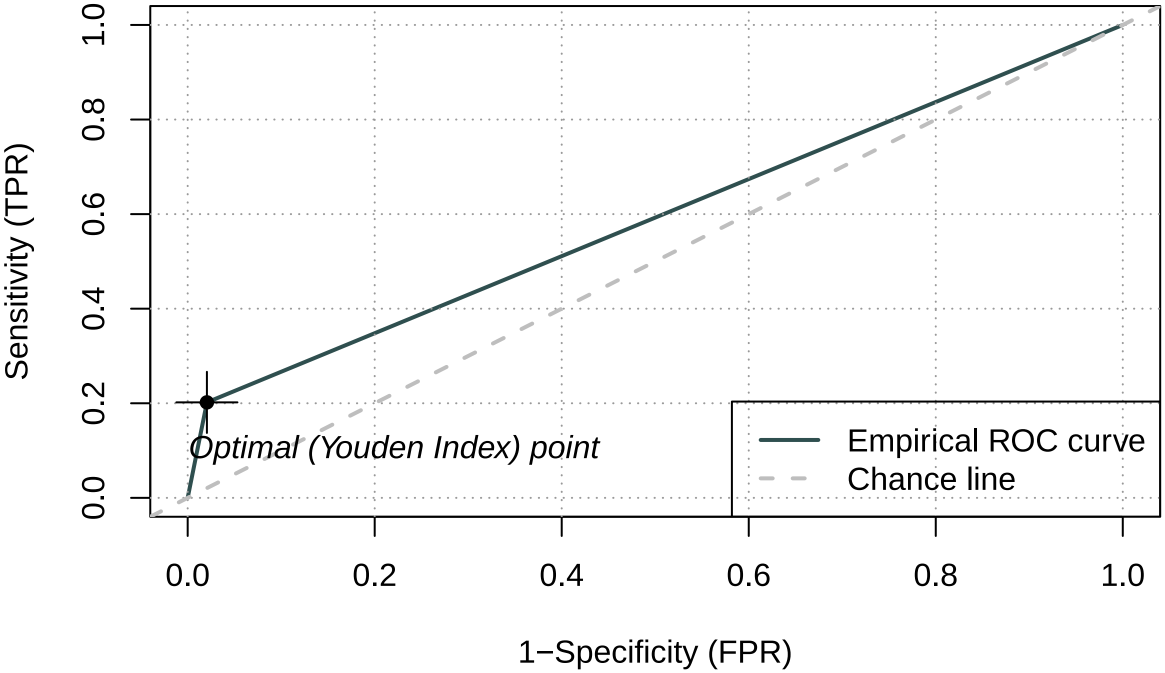
**Evaluation for Logistic Model**

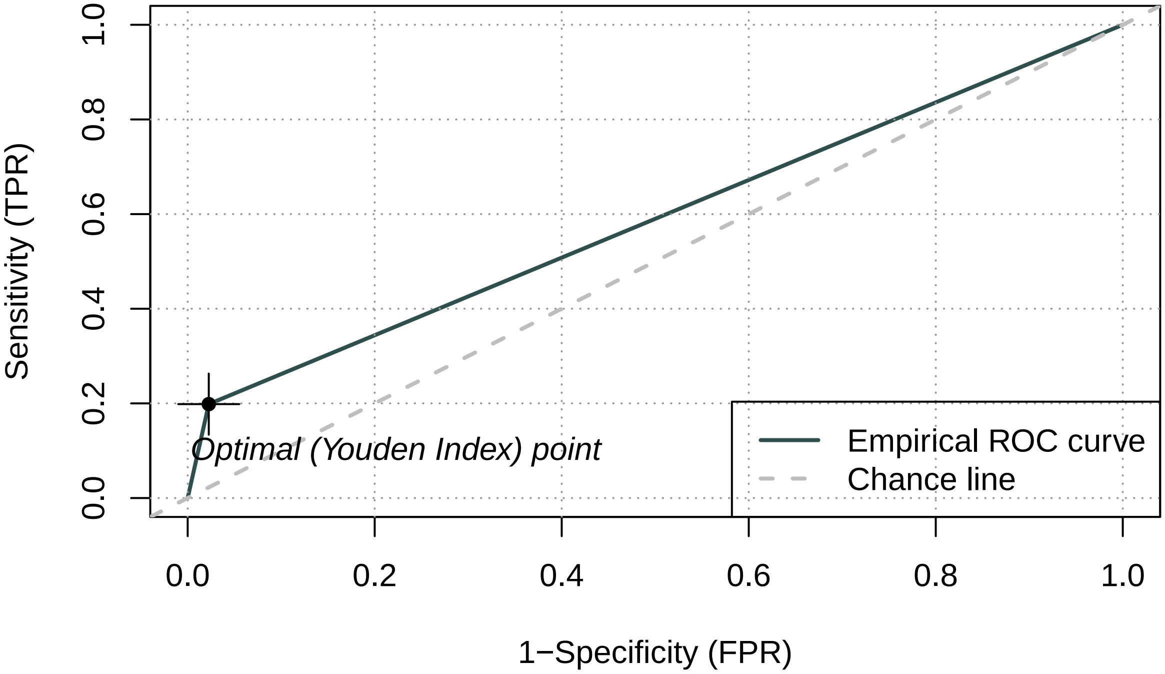
This is a model which we build using the logistics regression classification method. Using the variables that we selected on from earlier, which are Marital, Age, Amt_credit, PayRec_Sep, PayRec_Aug and PayRec_Jul as the variables, our results for the model is as follows:
Logistics regression Forward Backward with Y ~ Marital + Age + Amt_credit + PayRec_Sep + PayRec_Aug + PayRec_Jul:
Test Accuracy: 0.708
Train AUC: 0.6629
Test AUC: 0.5036

Test Harmonic Mean: 0.237

From the results, although the observed accuracy is quite high at 0.708, the harmonic mean is quite low at 0.237, which suggests a possible issue with the actual performance of our model. Furthermore, the AUC in test dataset is very low, which indicates that the model does not perform better by a significant margin relative to random selection.

### SVM

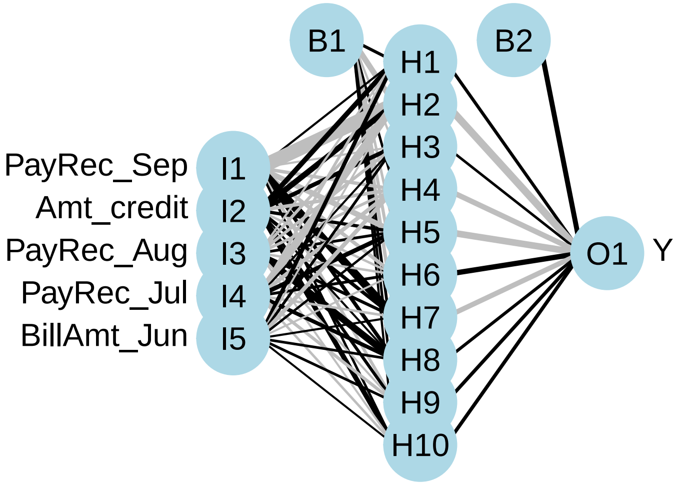


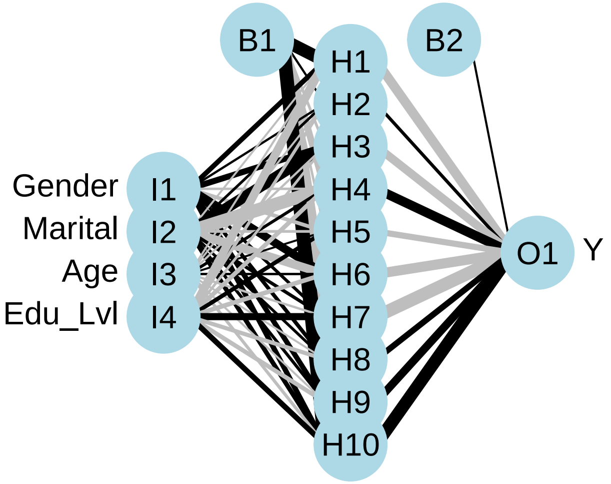


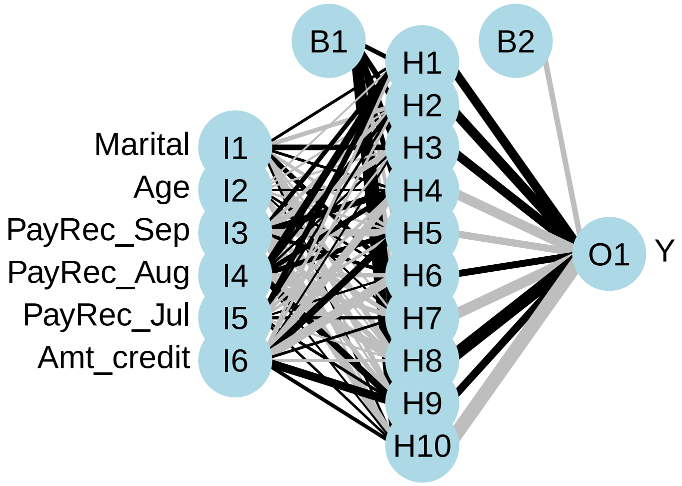
**Evaluation for SVM model**

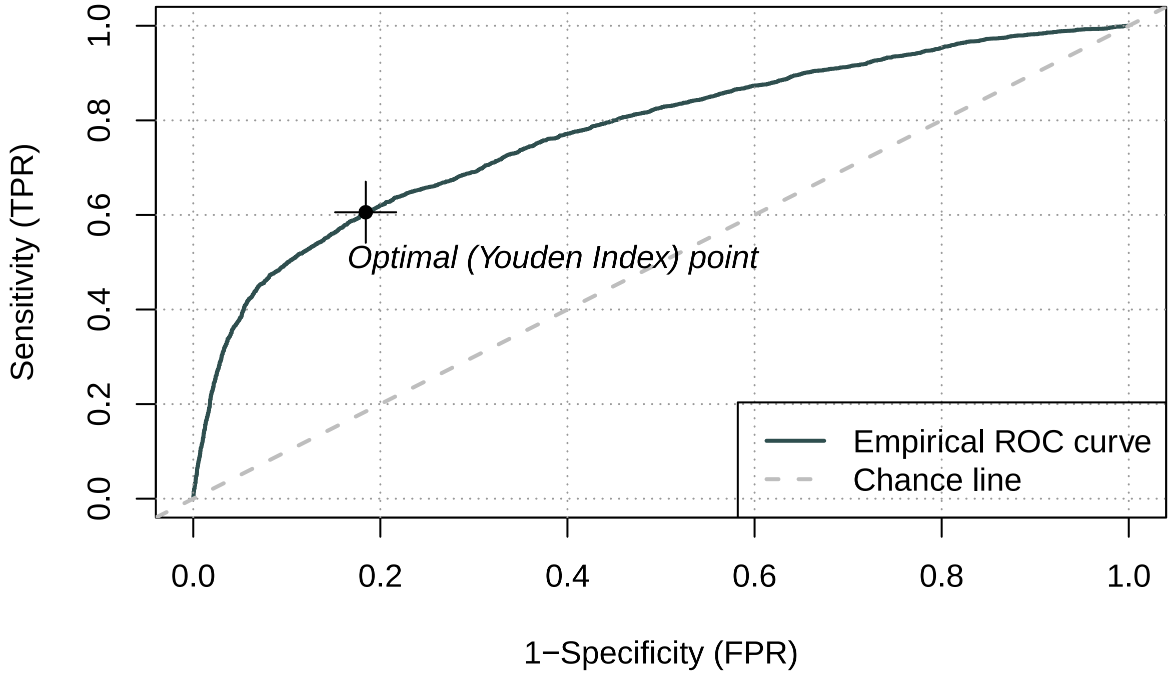
This is a model which we build using the support vector machine classification method. C-classfication, as well as radial kernel is used as our final settings for the model, based on our repeated testing on the test dataset. Using the variables that we selected on from earlier, which are Marital, Age, Amt_credit, PayRec_Sep, PayRec_Aug and PayRec_Jul as the variables, our results for the model is as follows:
SVM regression Forward Backward with Y ~ Marital + Age + Amt_credit + PayRec_Sep + PayRec_Aug + PayRec_Jul:
Test Accuracy: 0.805
Train AUC: 0.591
Test AUC: 0.588
Test Harmonic Mean: 0.330
From the results above, although the observed accuracy is quite high as well at 0.805, the harmonic mean for our test dataset is again quite low at 0.330, which suggests a possible issue with the actual performance of our model. However, from a AUC review, the train and test datasets achieved better results than that of the logistics regression model.

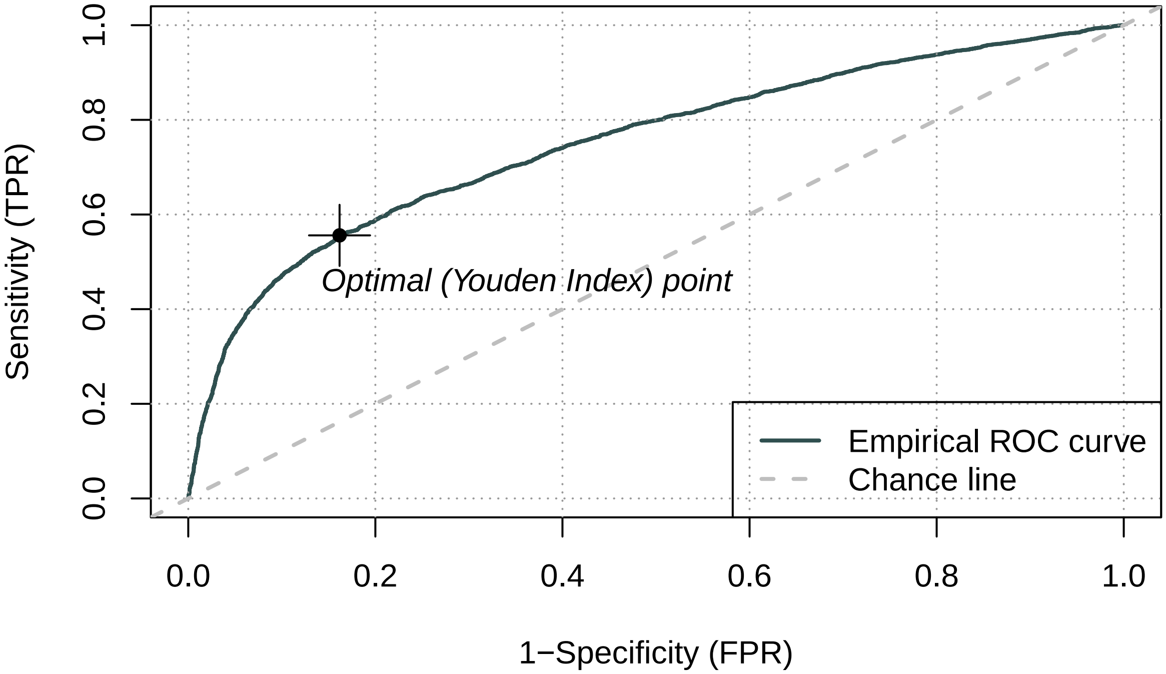
## Neural Network






B1
B2
Marital
Age
PayRec_Sep
PayRec_Aug
PayRec_Jul
Amt_credit
I1
I2
I3
I4
I5
I6
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
O1
Y


Sensitivity (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
1−Specificity (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Optimal (Youden Index) point
Empirical ROC curve
Chance line



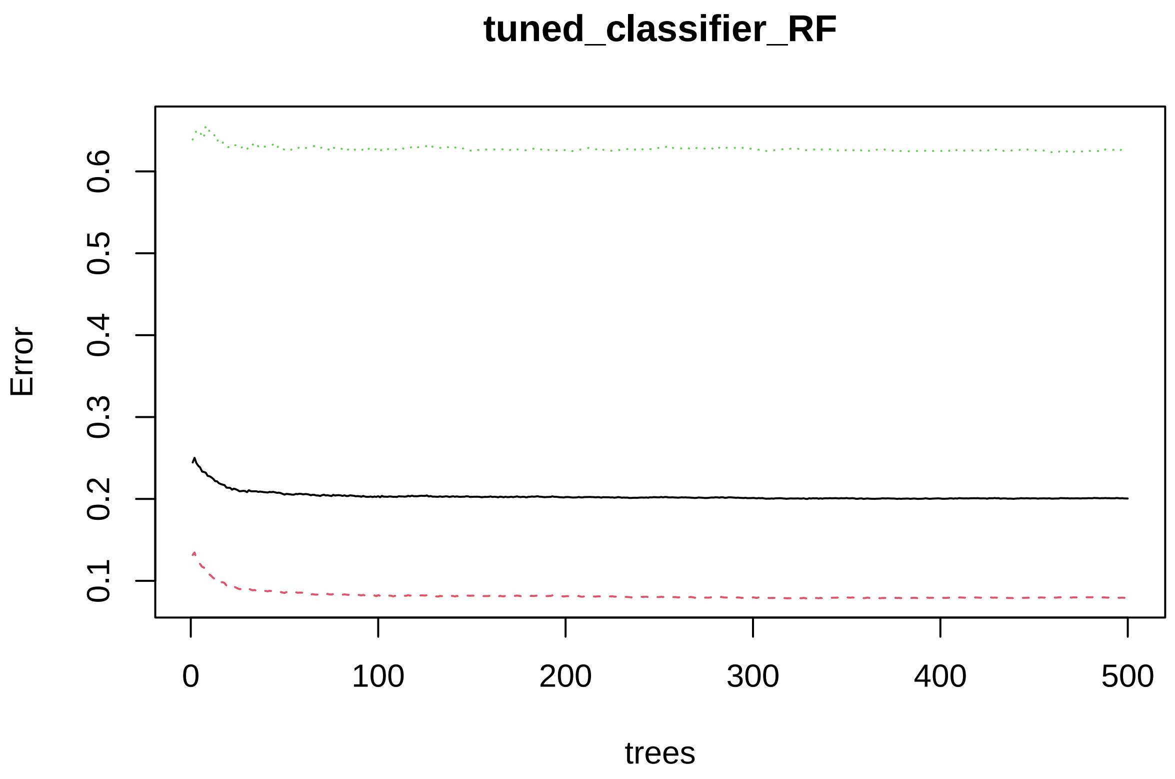
**Evaluation for Neural Network model**

This is a model which we build using the neural network classification method. A value of 1000 as max iterations to run, 10 hidden nodes and a decay factor of 0.08 is set for the model to prevent overfitting and achieve the best results. Using the variables that we selected on from earlier, which are Marital, Age, Amt_credit, PayRec_Sep, PayRec_Aug and PayRec_Jul as the variables, our results for the model is as follows:
Neural Network Forward Backward with Y ~ Marital + Age + Amt_credit + PayRec_Sep + PayRec_Aug + PayRec_Jul:
Test Accuracy: 0.818
Train AUC: 0.653
Test AUC: 0.647
Test Harmonic Mean: 0.501
From the results above, as we can see, the test accuracy is again quite high albeit still remaining quite close to the null accuracy. We can see that the harmonic mean still remains low, and it suggests that the model may not perform as good as we think it is. For AUC, the AUC for test and train set are close, which suggests overfitting is not an issue. However, we still believe that we can create a better model make the predictions.

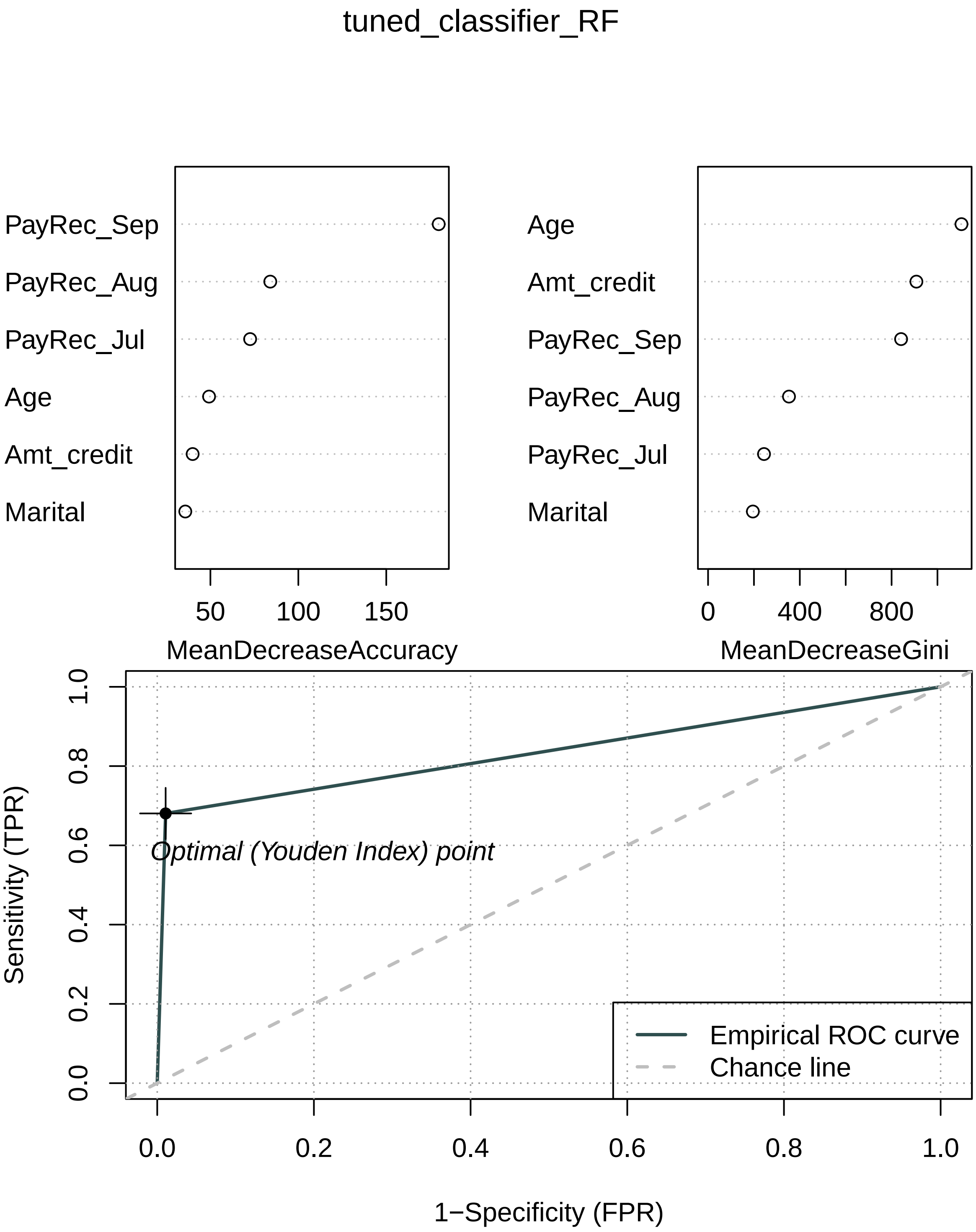
## Random Forest Classifier

In this case, mtry = 4 is the best mtry as it has least OOB error. Coincidentally, mtry = 4 was also used as default mtry.

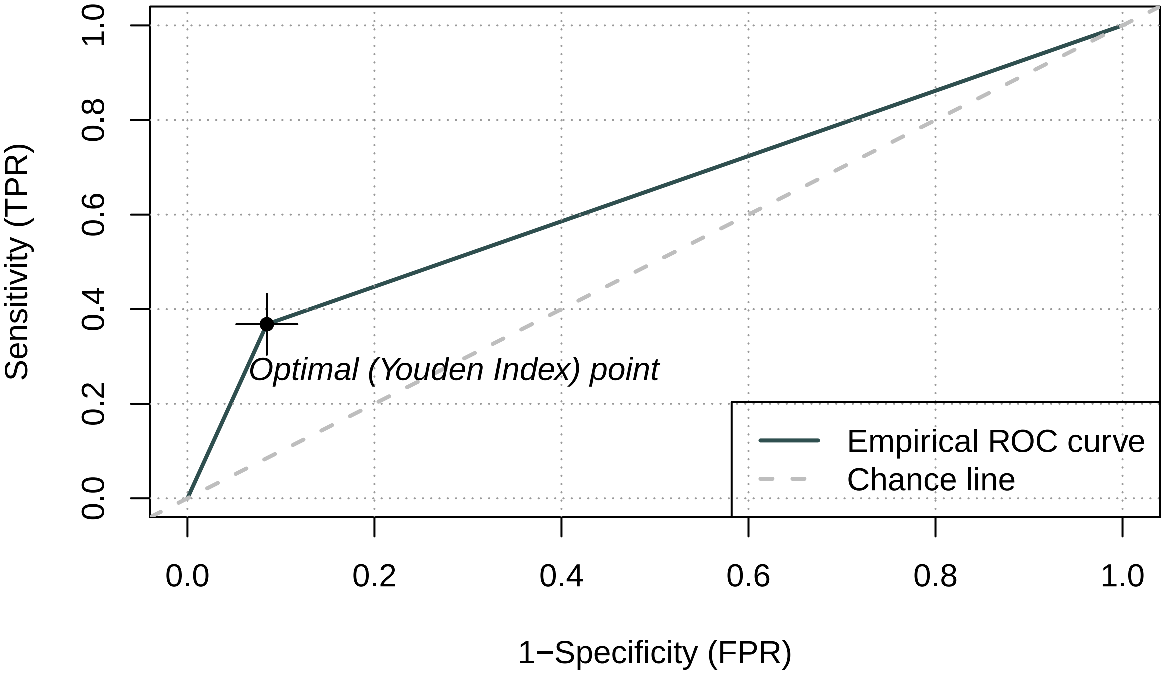
```
##    y_pred
##         0     1
##   0 10507   976
##   1  2065  1200
```



```
##                    0         1 MeanDecreaseAccuracy MeanDecreaseGini
## Marital     40.89120 -3.973139             35.72078         195.3995
## Age         49.26334  4.949118             49.23292        1104.6684
## Amt_credit  33.85993 16.085824             39.90170         908.1640
## PayRec_Sep 131.88687 69.726807            179.82329         841.4174
## PayRec_Aug  81.05604 -7.127203             84.04086         352.7161
## PayRec_Jul  54.28743 37.671283             72.59057         243.8465
```



tuned_classifier_RF
PayRec_Sep
PayRec_Aug
PayRec_Jul
Age
Amt_credit
Marital
50
100
150
MeanDecreaseAccuracy
Age
Amt_credit
PayRec_Sep
PayRec_Aug
PayRec_Jul
Marital
0
400
800
MeanDecreaseGini
Sensitivity (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
Optimal (Youden Index) point
Empirical ROC curve
Chance line
0.0
0.2
0.4
0.6
0.8
1.0
1−Specificity (FPR)



**Evaluation for Random Forest model**

This is a model which we build using the random forest decision tree classification method. From our repeated training, the number of nodes is set at 4, with TRUE set for gini importance and number of trees at 500. Using the variables that we selected on from earlier, which are Marital, Age, Amt_credit, PayRec_Sep, PayRec_Aug and PayRec_Jul as the variables, our results for the model is as follows:

**Seed 120**

**Random Forest with all variables:**

Accuracy: 0.861
Train AUC: 1
Test AUC: 0.6241

**Random Forest Chi Square with Y ~ Edu_Lvl + PayRec_May + PayRec_Jun + PayRec_Jul + PayRec_Aug:**

Accuracy: 0.846
Train AUC: 0.5633
Test AUC: 0.5413

**Random Forest first 5 of Boruta with Y ~ PayRec_Sep + PayRec_Aug + PayRec_Jul + BillAmt_Jun + PayRec_Jun:**

Accuracy: 0.845
Train AUC: 0.753
Test AUC: 0.618

**Seed 2**

**Random Forest Forward Backward with Y ~ Marital + Age + Amt_credit + PayRec_Sep + PayRec_Aug + PayRec_Jul:**

Accuracy: 0.795

Train AUC: 0.832
Test AUC: 0.642

## Final Review and Discussion

As we run the model using Marital + Age + Amt_credit + PayRec_Sep + PayRec_Aug + PayRec_Jul, we observed the train AUC & test AUC values closer to 1. A model which has AUC near to the 1 which means it has a good measure of separability. By analogy, the higher the AUC, the better the model is at forecasting clients who have tendencies to default on the bank.

Overall, from all the models we run as shown above, despite the increasing harmonic means and AUCs we achieved across models, we are still unable to achieve a good prediction score for all the models.

From our repeated model trainings, we realise that taking a huge number of variables in as predictors does not only makes the model more complex, but also may overfit to the test dataset. On the other hand, too few predictors may lead to a model that would not classfiy the data well. Therefore, based on the importance of each variable, we decided to come up with the variables as stated above as our best features to predict the model.

This is due to a few factors, namely the small number of predictors we selected for our models to achieve more efficient and quick predictions without wastage of resources, potential overfitting issues of using too much predictors as shown in our random forest run for a model with all the variables included, as well as some potential data integrity issues that arises from the unexplained values that are found in our data visualisation. Unbalancd nature of our dataset also suggests that there may be harder to predict due to the lower number of default clients as well.

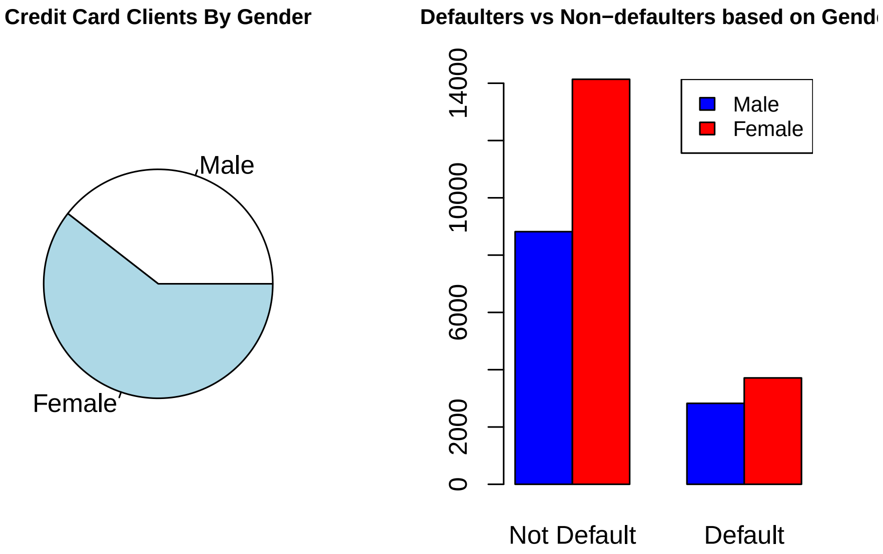
However, there is no denial that our model is not perfect and there are ways to improve on it. For example, the lower score for both AUC, as well as harmonic means suggests that there are still ways to improve on the model. We believe that with better knowledge and understanding of the data, we will be able to create a better model, as compared to the satisfactory model we have now.

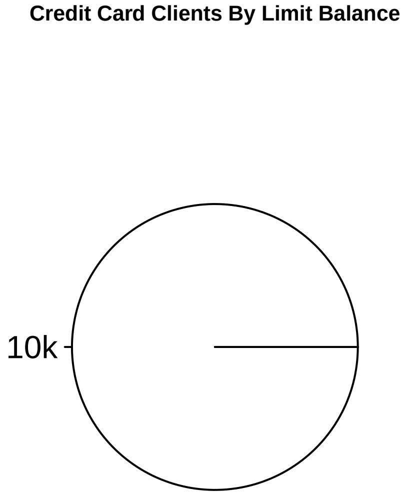
While reviewing our project, we discovered that the dataset is unbalanced which could result in bias in the model. This can be shown though the figures below.

### Breakdown of Customers based on default status

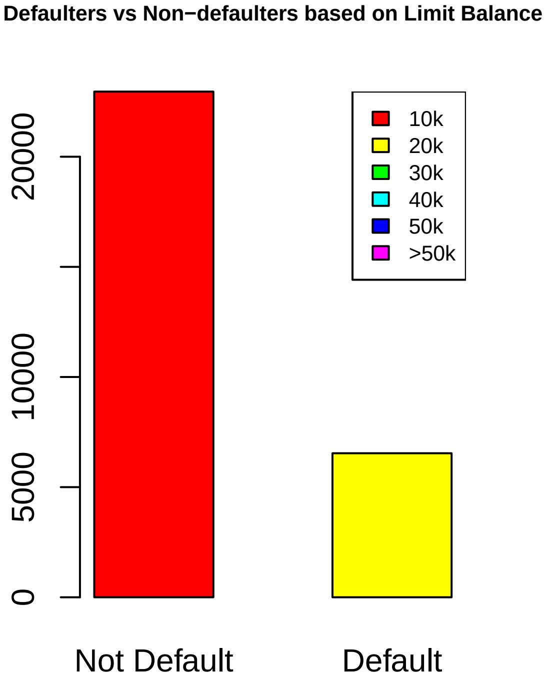
```
## Not Default     Default
##       22957        6539
```

### Breakdown of Customers based on Gender and Limit Balance





Credit Card Clients By Limit Balance



Defaulters vs Non−defaulters based on Limit Balance