BUAN 6357: Advanced Business Analytics using R

Project Report

Credit Card Default Prediction

Adityan Rajendran
(AXR180073)

Table of Contents

Executive Summary …… 1

Project Background …… 2

Data Selection …… 3

BI model …… 5

Conclusion …… 26

Reference …… 27

Executive Summary

In the year of 1990, the Taiwanese government allowed the formation of new banks. The "new" banks, with a goal of expanding their business and multiplying profits, lent large sums of money to real estate companies. They also took to newer businesses such as – *credit cards* and cash cards. The banks went all ends to encourage people to apply for credit cards such as lowering the requirements, targeting young people and running expensive commercials. However, after a few years of expansion, in February 2006, debt from credit cards reached $268 billion (USD). More than half a million people became "Credit card slaves", a term coined in Taiwan to refer to people who could pay only the minimum balance on their credit card debt. Therefore, this project aims to awake our awareness about such a financial crisis and help detect whether a credit card customer will default on his payment or not.

The dataset that is chosen was originally donated to the 'Center for Machine Learning and Intelligent Systems' at University of California, Irvine for research aimed at the case of customers' default payments in Taiwan to determine the predictive accuracy of probability of default for various data mining algorithms. This dataset consists credit data of 30000 customers. It comprises of 23 predictor variables and 1 target variable, '*default.payment.next.month*'. The predictor variables include demographic variables such as Sex, Education, Gender, Age etc. and credit-based metrics such as Payment status, Amount paid and Bill amount.

We used R, a statistical computing and graphics tool for building our BI model. At first, Exploratory Data Analysis was performed to understand and visualize the dataset and to develop a general idea for further analysis. Several correlation plots, histograms and bar charts were plotted to uncover the relationship and distribution among variables, for ex: University goers had the highest default rate, Customers over the age of 60 had the lowest default rate etc.

Then, after some preliminary data cleaning and pre-processing, several classification models were built, including Decision Tree, Support Vector Machines, Random Forest and Logistic regression, to predict whether a given credit card client, for the next month, would default on his payment or not. The data was sampled with 70% for training and 30% for validation, post-which the models were built and visualized with required plots. The models were then evaluated using confusion matrix and ROC.

After performing the evaluation for all the models, it was concluded that the Logistic Regression model was the best model with the best ROC metric, and it could be used for obtaining predictions on payment default for credit card customers.

Project Background

In the early 1990s, Taiwan was experiencing Financial Liberalization which lead to the number of banks becoming doubled in an already crowded market. During this time, new and less experienced banks entered the financial markets which intensified the competition among banks for increased market share, leading to relaxed lending standards and stronger credit expansion. With card-issuing banks over-issuing cash and credit cards to unqualified applicants and at the same time, with most of the cardholders over-using credit cards irrespective of their repayment ability, this led to a Credit card debt crisis in 2005 with delinquency expected to peak in the third quarter of 2006.

The credit cardholders, whose information is captured in this dataset, have a high rate of defaulting their payments in the month of October 2005, a rate equal to almost 25%. While a lot of research quotes the Financial Liberalization as the core reason for such defaulting, it is only one factor that has caused a customer to default. Information on other demographic and payment factors relating to the customer is necessary. Though the crisis occurred across Taiwan entirely affecting millions of customers, a small subset of the entire population base is taken and studied for various other reasons of default and possible predictions that are accurate enough to foresee defaults for a particular customer profile.

Hence, this project is proposed to be built to study the demographic, financial factors and other diagnostic variables of Credit card customers in the timeline of April to September 2005 and develop a classification model which could predict the default of payment among the customers given a set of demographic & financial information.

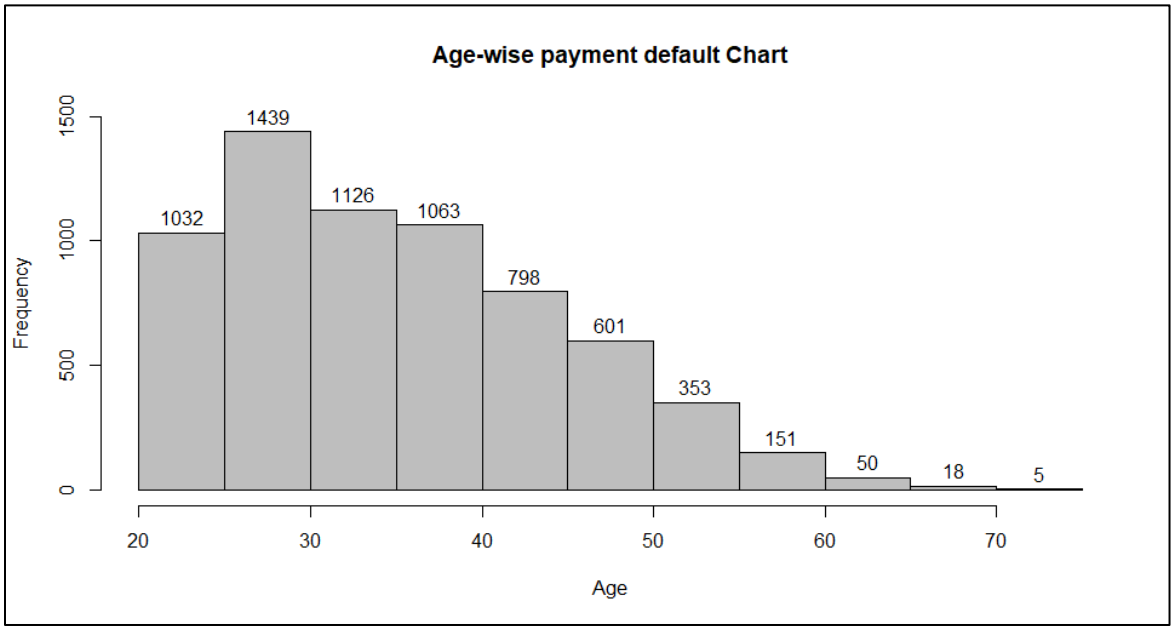
Data Selection

The dataset that is chosen is obtained from Kaggle (www.kaggle.com), donated by the 'Center for Machine Learning and Intelligent Systems' at University of California, Irvine. It is named as 'Default of Credit Card Clients Dataset'.

It comprises of a typical customer-centric payment and demographic profile of 30000 credit cardholders. Focusing on an age centric perspective, the number of customers in the age group of 25-35 were 12,938 in number and among them 2565 are identified as payment defaulters, which is an alarming rate of 20%.

The customer age range is given below:

Age	Number of Customers
20 -25	3,871
25-30	7,142
30-35	5,796
35-40	4,917
40-45	3,605
45-50	2,400
50-55	1,425
55-60	572
>60	272



The dataset consists of several predictor variables - demographic variables such as Age, Gender, Marital Status etc., payment related variables such as Payment status, Bill Amount, Paid amount for 6 months (April – September 2005) and one target variable – '*default.payment.next.month*'.

Approaching the dataset in a technical perspective, the dataset is in a 'Comma Separated' file – 'UCI_Credit_Card.csv'. We have credit cardholders with a minimum age of 21. The list of independent and target variables is listed below with a brief description. There are 23 independent variables and 1 target variable. The dataset has 30000 rows.

Independent Variables

Independent Variables	Description
ID	ID of the client
LIMIT_BAL	Amount of given credit in NT$ (New Taiwan Dollar)
SEX	Gender (1 = male, 2 = female)
EDUCATION	Client's education level (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6	Repayment status for months from April – September 2005 (1=payment delay for 1 month, 2=payment delay for 2 months……)
BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6	Amount of bill statements for months April – September 2005 (in NT$)
PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	Amount of previous payments for months April – September 2005 (in NT$)

Target/Dependent Variable

Target/Dependent Variable	Description
default.payment.next.month	Default payment (1=yes, 0=no) (6636 counts for 1, 23364 counts for 0)

There are 10 categorical variables (excluding the output variable) and 13 numerical variables. The ID variable is excluded from the analysis purview.

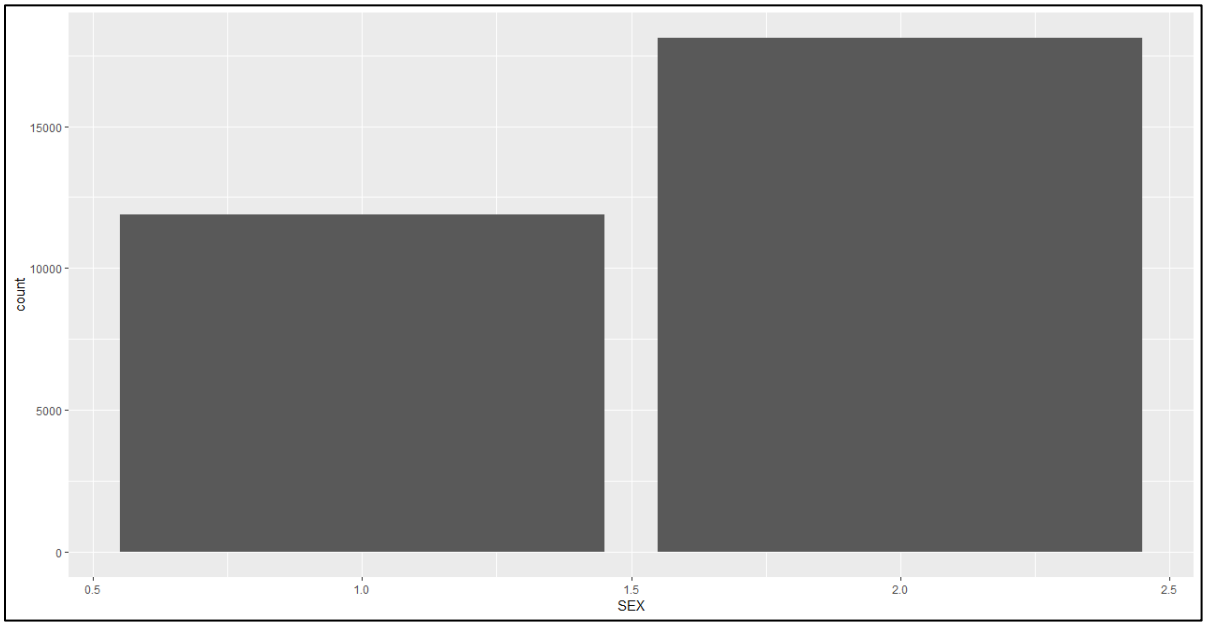
BI Model

Data Pre-processing and Visualization

Firstly, we present data in a form that makes sense to us so that we can have a general idea about the data and find directions for further analysis.

Since there are a mix of categorical and numerical variables present in the dataset, we start by visualizing the categorical variables to get a sense of distribution and correctness. The data pre-processing is initiated with the check for any missing/NULL values and presence of unusual number of zeroes. It is concluded that there is neither any missing or NULL values nor a presence of large number of zeroes in any of the features.

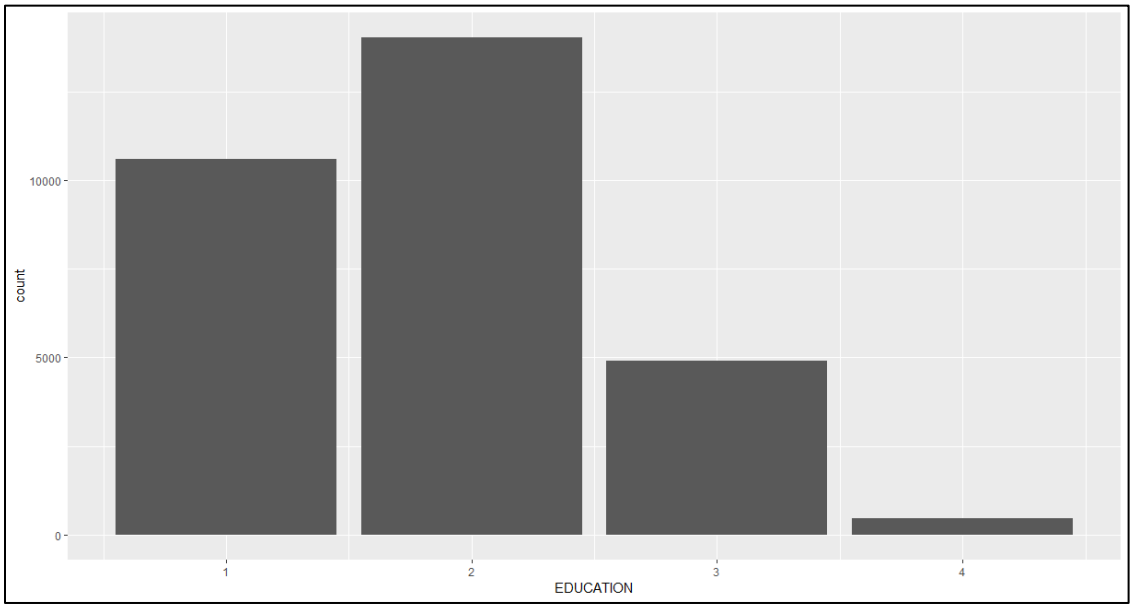
All the input variables whose visualization is realized below are subject to certain amount of data cleaning and pre-processing. The type of cleaning/processing is explicitly mentioned along with each plot (if done on that variable).
The bar plots for the categorical variables are posted below:

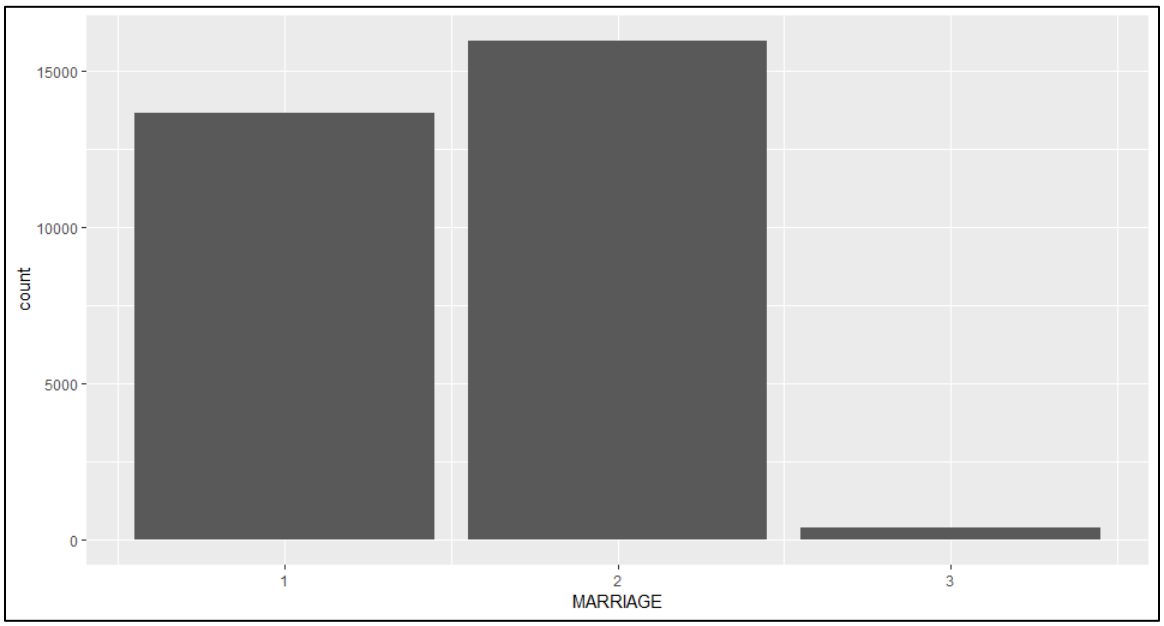


This bar plot of the SEX variable shows us that the number of females is higher than the number of males. (Males (1) – 40%, Females (2) – 60%)

The bar plot of EDUCATION variable tells us the distribution of credit cardholders based on their education. We have 35% attending graduate school, 45% attending university, 16% in high school and the rest have various other education levels*. For the EDUCATION variable, there were undocumented factor levels 5 and 6 which were merged to 4 for simplicity and aptness.*

The MARRIAGE variable distribution tells us that most of the customers are Single – 53% with married people constituting 45% and the rest have various other marital statuses.

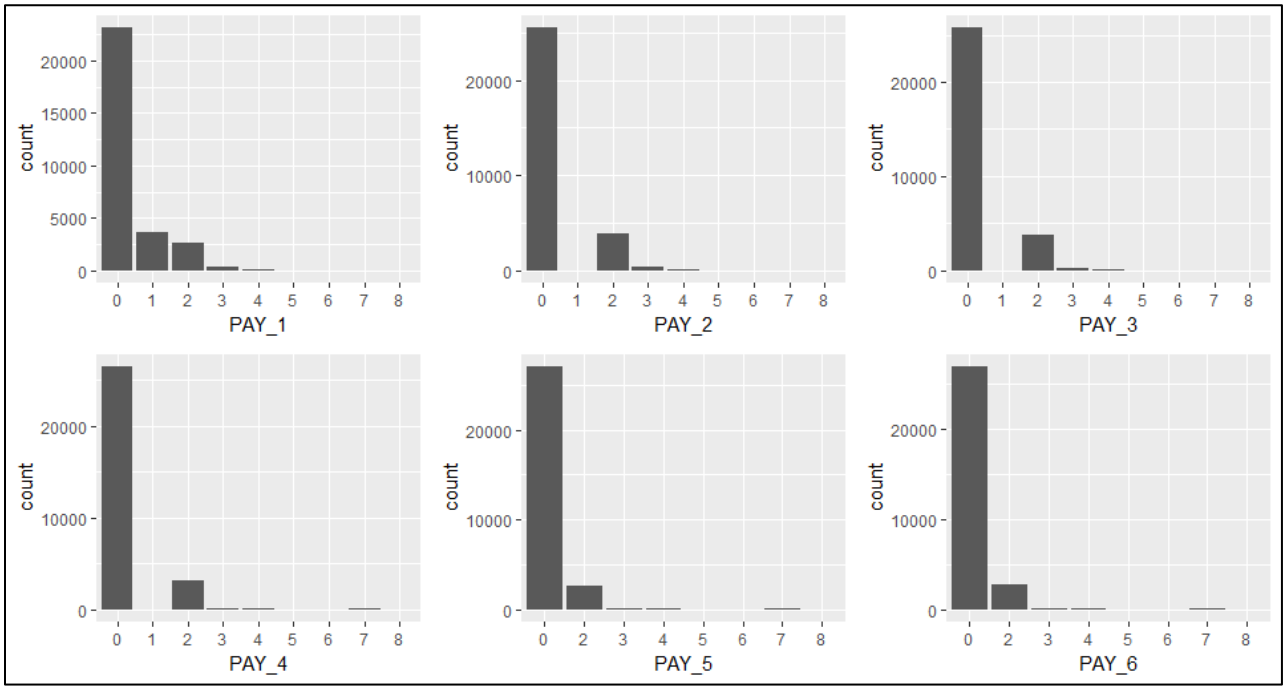




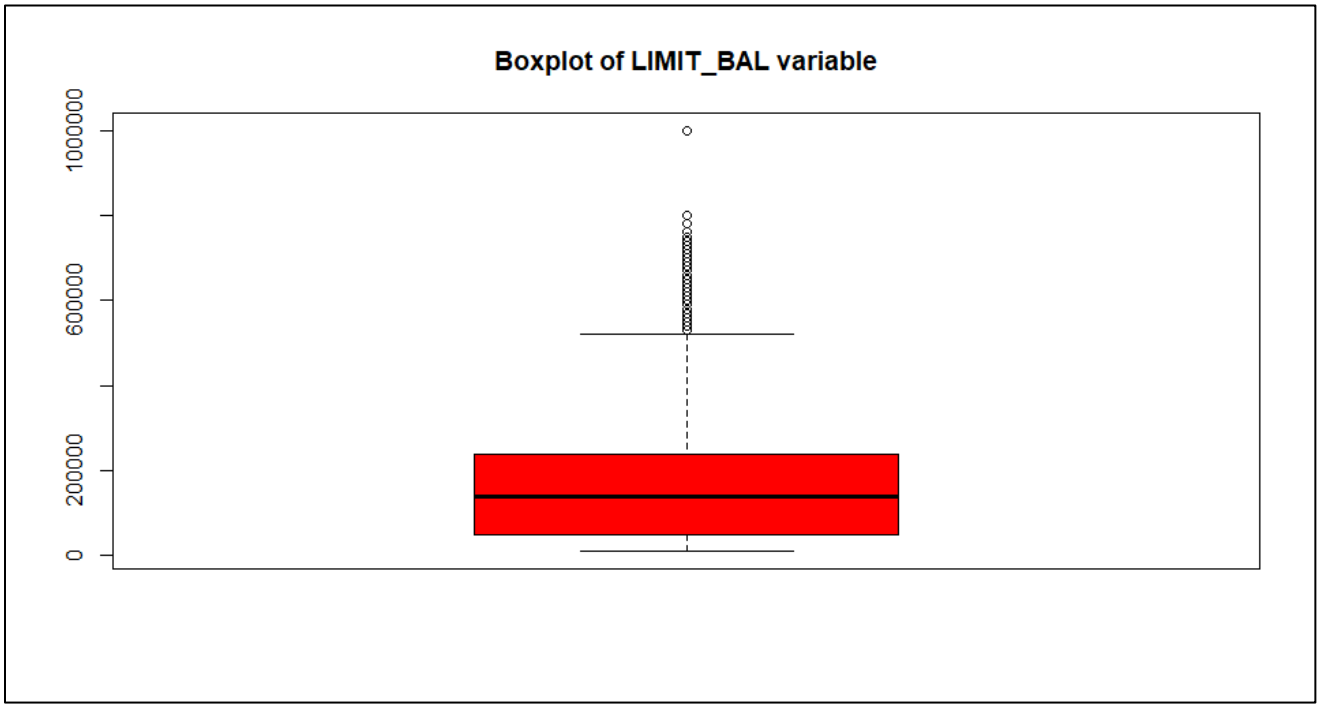
We have 6 more categorical variables representing the payment status with various levels representing payment delay in months. Ex: PAY_2 represents the payment status for the month of August 2005and a value of PAY_2 <- 4 means that the payment is due for 4 months. The plots are provided below:

The distributions show that most of the credit cardholders are not defaulting since there is a spike at zero. Most of the customers defaulting, default with a due of 1 or 2 months with a minor amount defaulting with a due greater than 2 months.

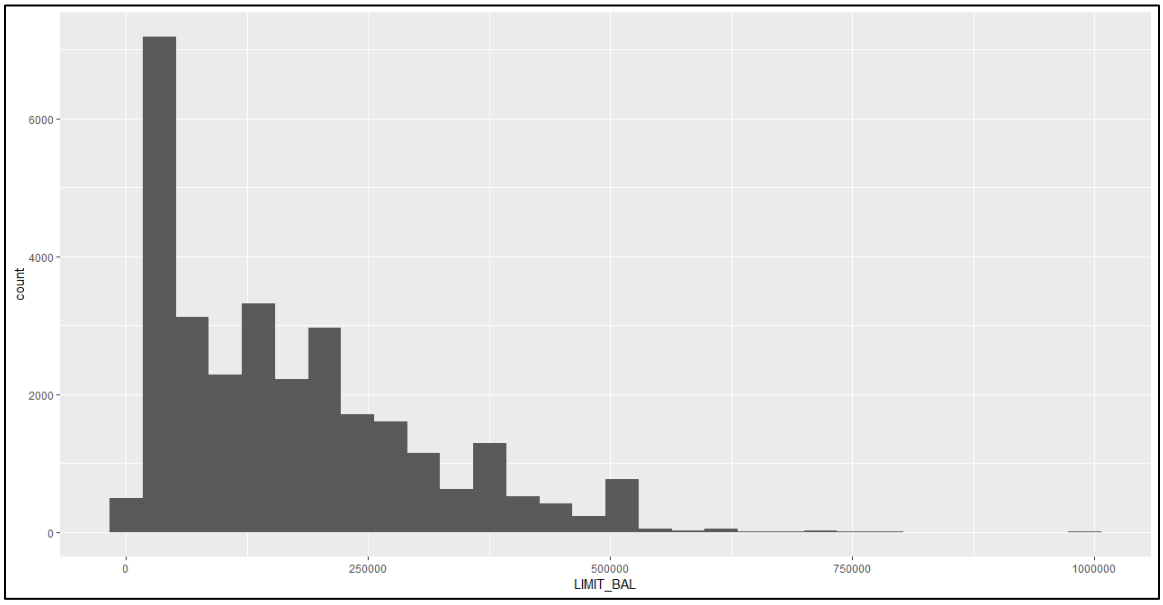
For the PAY_X variables, there were payment statuses with negative values representing credit and they were merged to 0 for simplicity and aptness. Zero indicates that payment is done duly within the deadline.



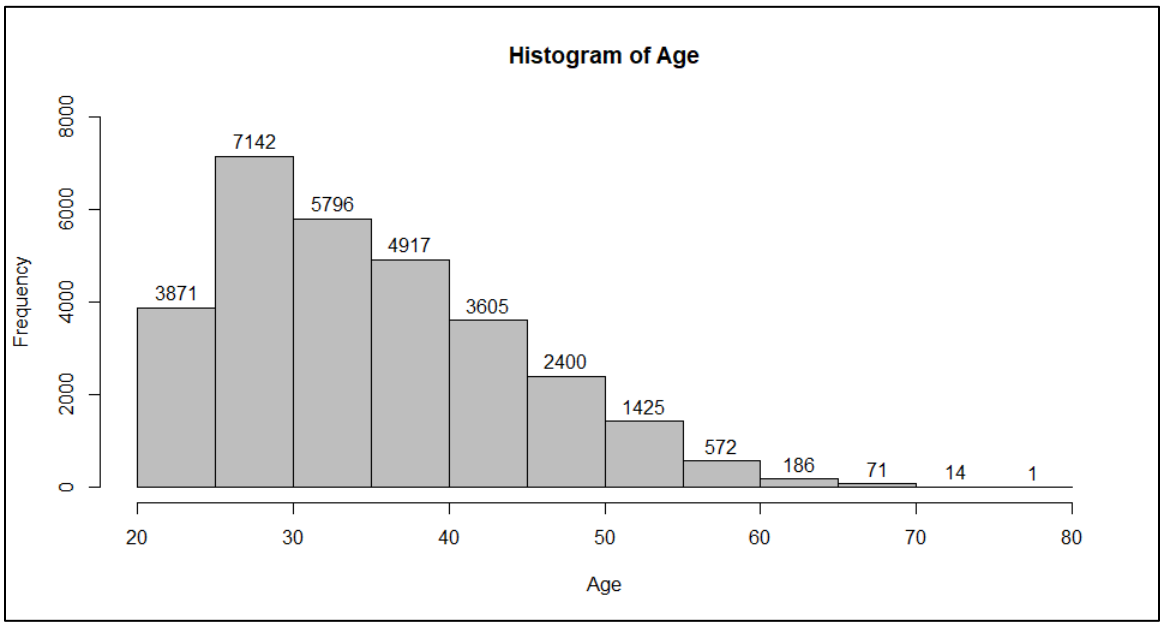
Having visualized the categorical variables, we move to the numerical variables.
The plots are shown below:



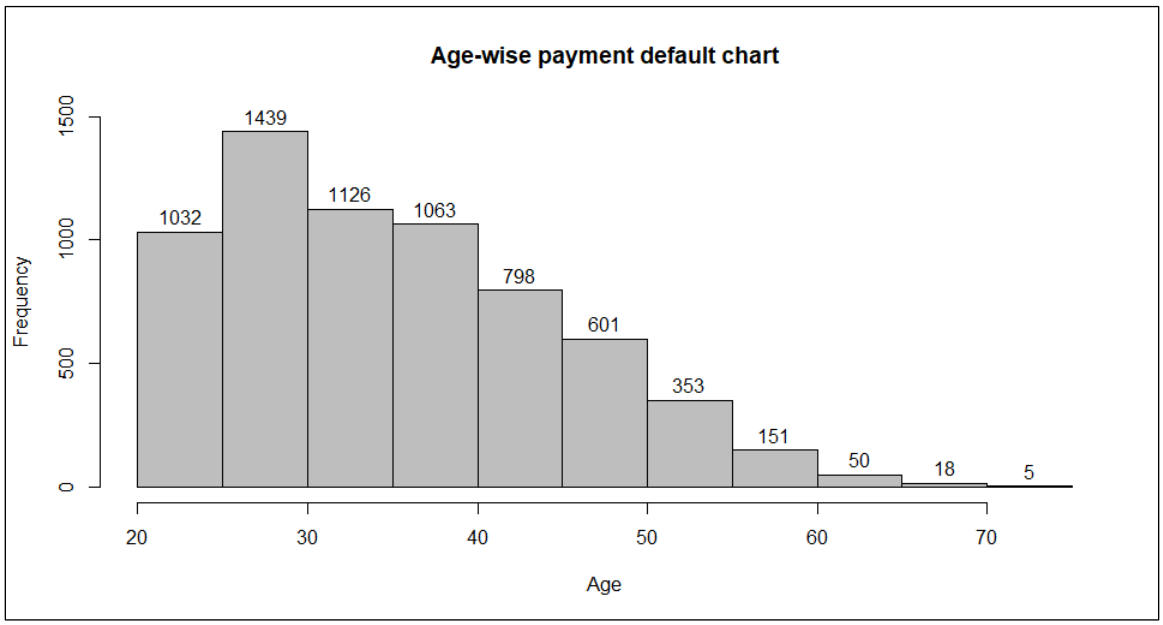
Visualizing the LIMIT_BAL variable, it represents the amount of outstanding credit in the credit cardholder's account. Looking at the boxplot above, we see that more than 75% of customers have an outstanding credit of less than 300,000 NT$, with the median being approximately 150,000 NT$. There is a considerable number of outliers whose credit is more than 500,000 NT$ and there is one strikingly odd customer with an outstanding credit of 1000,000 NT$. The histogram is shown below for conformation.



The below shown histogram of AGE variable tells us that customers are majorly young with an age range of 25-30, followed by 30-35 and so on. It can be inferred that customers are primarily the younger workforce.



Offering an insightful perspective using the AGE variable is the below plot showing age-wise payment default chart. We can see that customers in the age range of 25-30 default the most followed by 30-35 and so on., which reflects the free-spending mindset of the younger population base in that time (April – September 2005).



Proceeding to the set of BILL_AMT variables, which represent the monthly bill statements of customers from April – September 2005, summary statistics is employed to get a sense of the distributions (pasted below).

```
   BILL_AMT1          BILL_AMT2          BILL_AMT3
Min.   :-165580   Min.   :-69777   Min.   :-157264
1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666
Median :  22382   Median : 21200   Median :  20088
Mean   :  51223   Mean   : 49179   Mean   :  47013
3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165
Max.   : 964511   Max.   :983931   Max.   :1664089
   BILL_AMT4          BILL_AMT5          BILL_AMT6
Min.   :-170000   Min.   :-81334   Min.   :-339603
1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256
Median :  19052   Median : 18104   Median :  17071
Mean   :  43263   Mean   : 40311   Mean   :  38872
3rd Qu.:  54506   3rd Qu.: 50190   3rd Qu.:  49198
Max.   : 891586   Max.   :927171   Max.   : 961664
```

It can be noted that monthly bill statements are widely distributed with a broad range including negative values which indicate credit pending to the customer.
Similarly, the summary statistics of the PAY_AMT variables is given below. The PAY_AMT indicates the amount of previously recorded payment in the given month. Ex: PAY_AMT1 represents the amount of payment made in the month previous to September 2005.

```
    PAY_AMT1          PAY_AMT2          PAY_AMT3
Min.   :     0   Min.   :       0   Min.   :     0
1st Qu.:  1000   1st Qu.:     833   1st Qu.:   390
Median :  2100   Median :    2009   Median :  1800
Mean   :  5664   Mean   :    5921   Mean   :  5226
3rd Qu.:  5006   3rd Qu.:    5000   3rd Qu.:  4505
Max.   :873552   Max.   :1684259   Max.   :896040
    PAY_AMT4          PAY_AMT5          PAY_AMT6
Min.   :     0   Min.   :     0   Min.   :     0
1st Qu.:   296   1st Qu.:   252   1st Qu.:   118
Median :  1500   Median :  1500   Median :  1500
Mean   :  4826   Mean   :  4799   Mean   :  5216
3rd Qu.:  4013   3rd Qu.:  4032   3rd Qu.:  4000
Max.   :621000   Max.   :426529   Max.   :528666
```

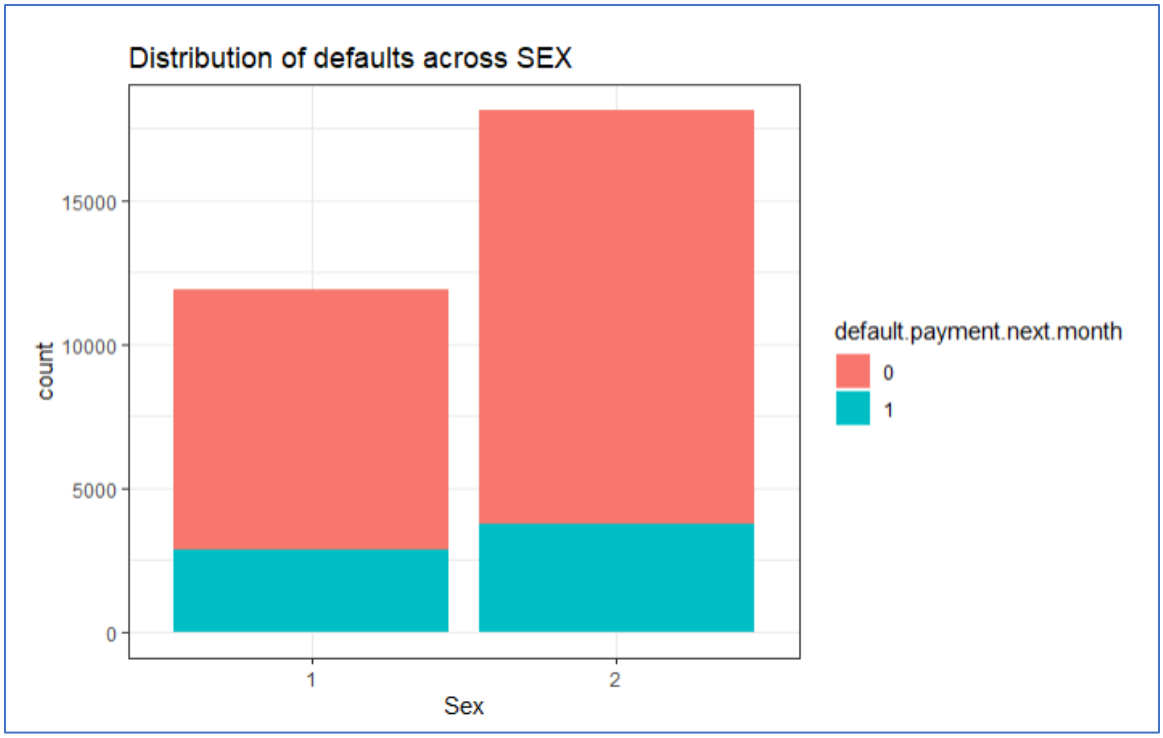
The payments range from zero to values up to 1,684,259 NT$.

Data Exploration and Analysis

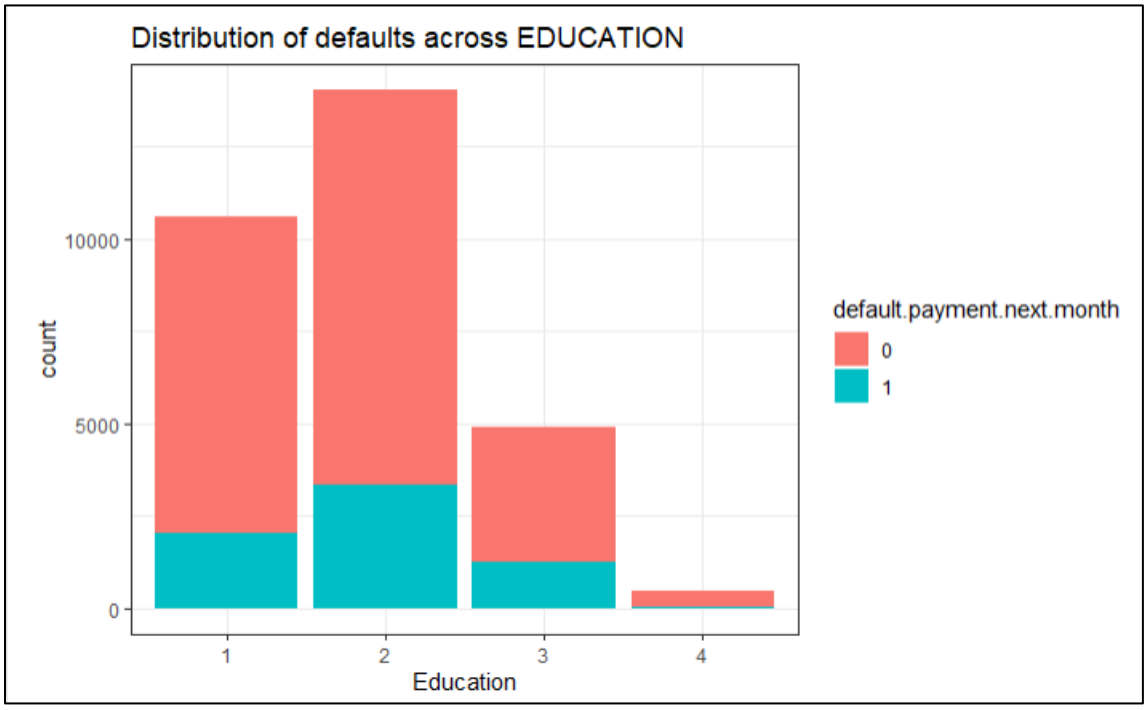
After visualizing the various categorical and numerical variables present in the dataset, a sense of their distribution is obtained.

We now try to understand how the input variables are distributed with respect to the target variable. A rough estimate can be made in determining the significant predictors for default payments and how each input variable varies based on the output or vice-versa.

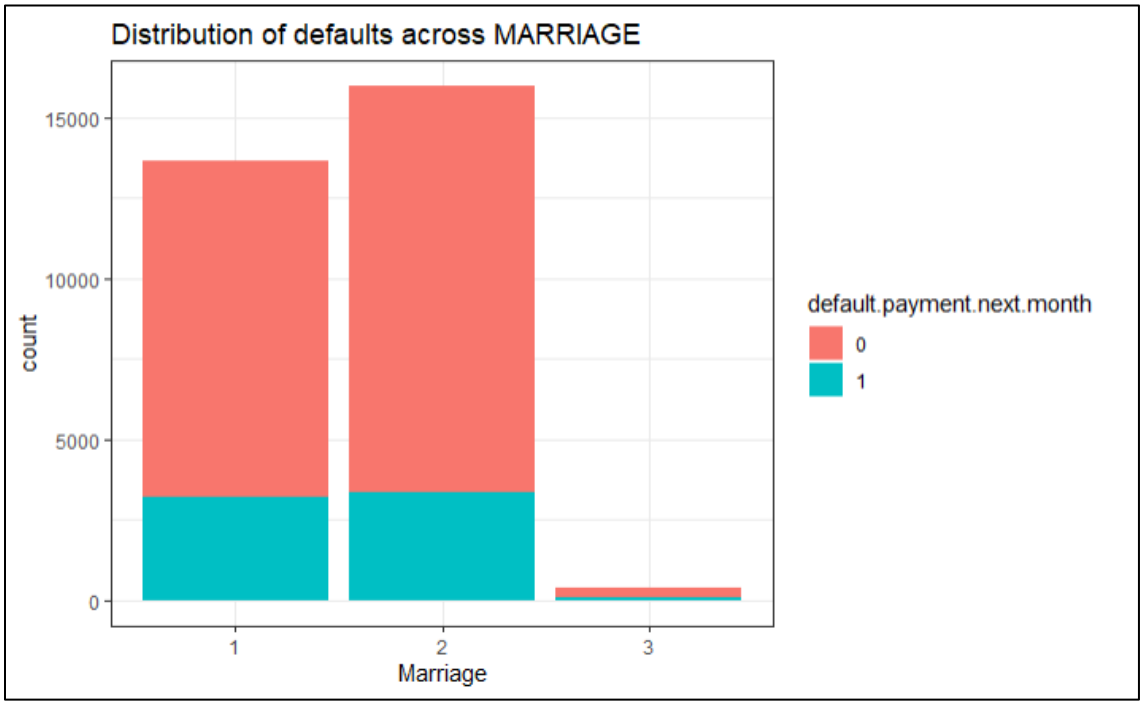
The below plot shows how each gender is jointly distributed in their constitution of the output variable. It can be noted that number of defaults is almost the same irrespective of the gender and gender size. Although, the customer base is predominantly female, the proportion of default is higher among males.



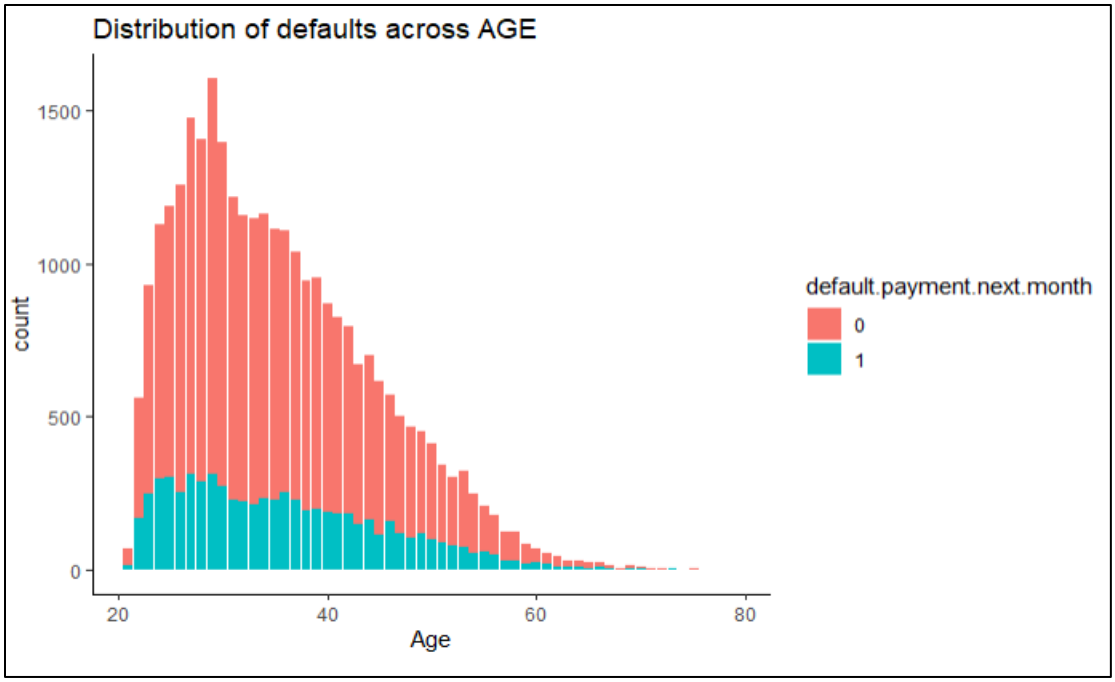
The subsequent plot below shows us the default rates for various categories in EDUCATION variable. It can be noted that large number of defaults occur among University (EDUCATION = 2) goers with a count of 3,330 out of 14,030. But, looking at a proportion, the highest default rate occurs among high schoolers (EDUCATION = 3) with a proportion of 25.16%.

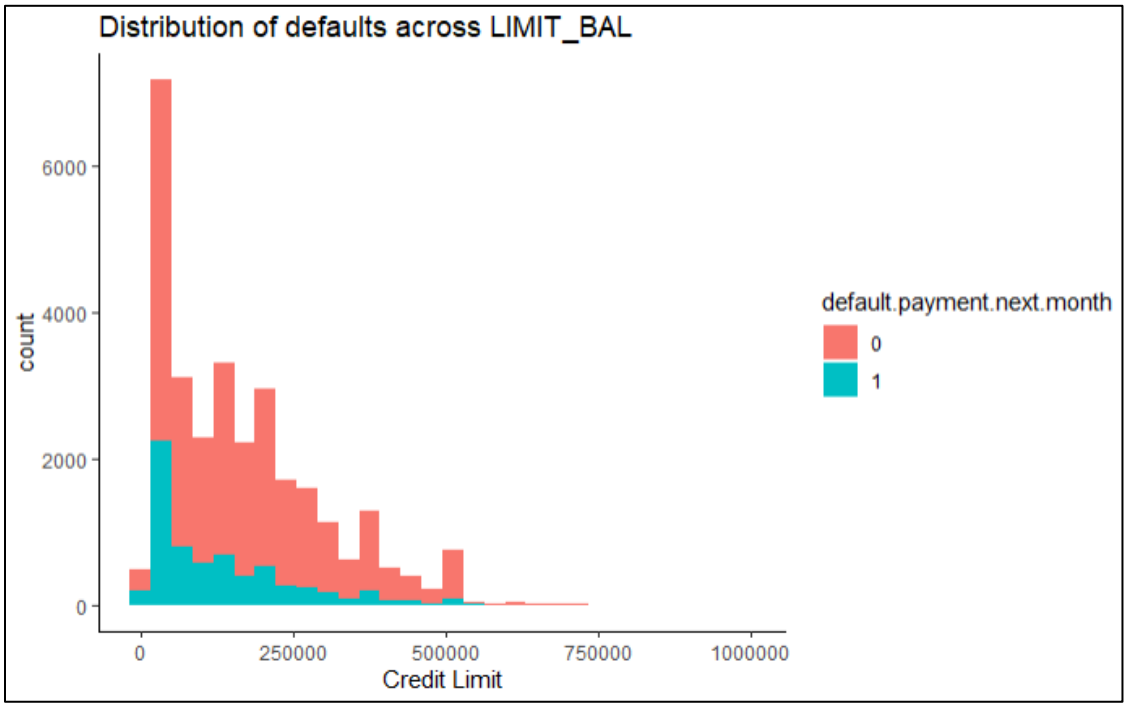


The distribution of defaults, with respect to various categories present in the MARRIAGE variable is studied using the below plot. It can be inferred that the number of defaulters is more or less the same for Single (MARRIAGE = 2) and Married (MARRIAGE = 1) customers. Most of the customers are Single while we have a meagre number for various other marital statuses. From a proportion perspective, Married customers have a high default rate equaling 23.47%.



Looking at the distribution plot of AGE Vs 'default.payment.next.month' below, it can be inferred that number of defaults decrease as AGE increases. This reflects the spending attitude of younger and older generations of people. Though Financial liberalization led to over-spending because of the over-use of credit cards, the most affected population base were the younger people in an age range of 25 – 45 years. It can also be noted that number of defaults among people in the age range of 20-25 is pretty less. There weren't many people in that age range who availed a credit card.

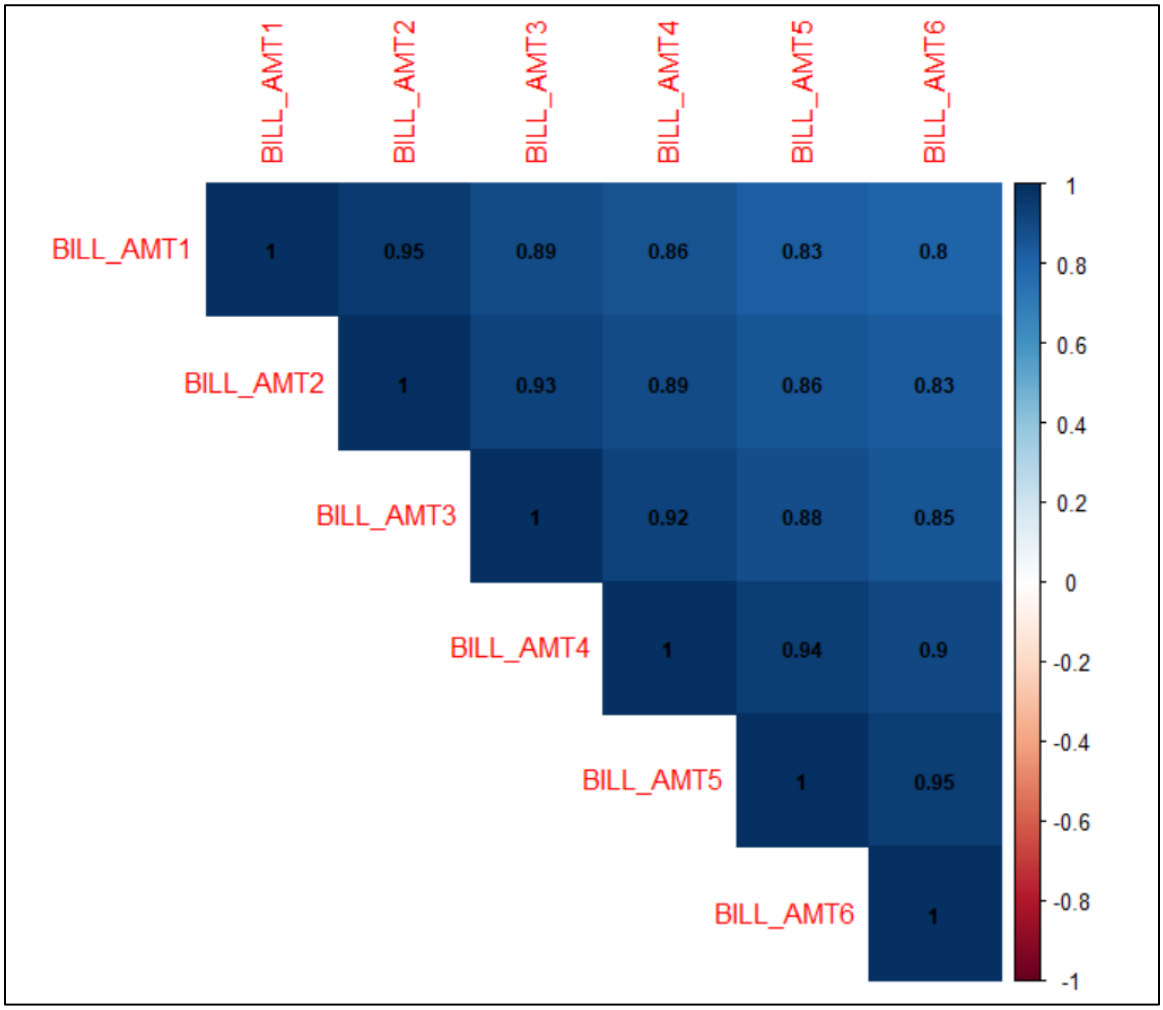




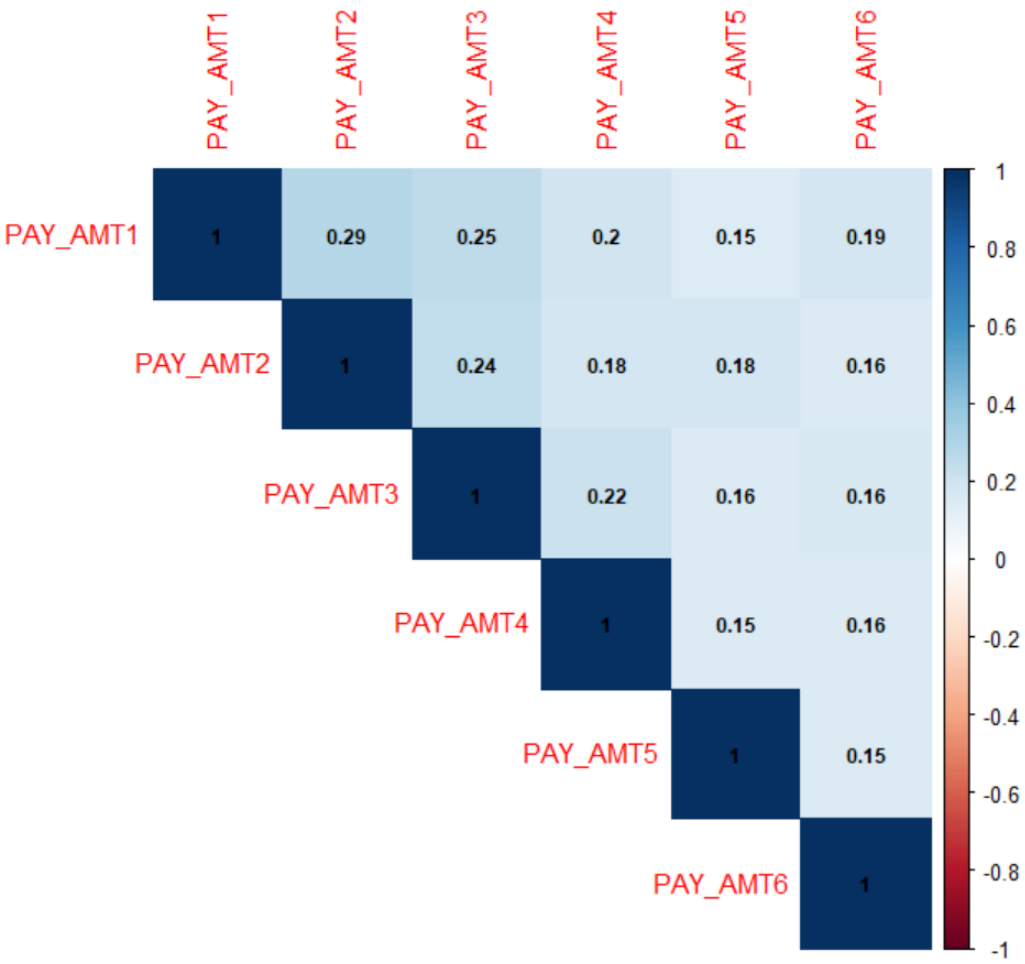
The above plot gives us a rough idea of how the outstanding credit balance of customers varies with payment default. It can be inferred that a majority of defaults happen when the outstanding credit balance is in the range of 31,250 – 125,000 NT$.

We compute the correlation plots for the BILL_AMT variables and the resulting correlation plot is pasted below. It is expected that a high correlation exists between the BILL_AMT variables because for a specific customer the usage is supposed to be more or

less the same every month. This is in fact confirmed by the correlation plot wherein we see a high correlation among the variables with the minimum being 0.8 and the maximum being 0.95. The correlations, however decrease as months progress.



The correlation plot for PAY_AMT variables is also computed and pasted below. No significant correlations exist among the variables telling us that the monthly payments were not consistent. They did not follow any given order as it was the time when customers were so hard hit by the financial crisis that they were unable to be regular in their earnings and payments.



Now that a broad sense of how variables are distributed and how they relate significantly to the target variable is obtained, we proceed to building a set of classification models to unearth the underlying pattern and significant predictor profiles which can best help us to predict the payment default for the next month.

Classification Models

The dataset has 23 independent variables and one target variable. We use this dataset to build classification models to predict whether a credit card customer, with a certain demographic and payment profile, will default or not on his next monthly payment.

1. Decision Tree

The first classification model for prediction is the decision tree model which is one of the simplest yet powerful model. The decision tree model leads to classification of the target variable based on 'rules' formed using independent variables.

The decision tree model, like every other tree algorithm is robust to missing values and outliers. Although the dataset does not contain any missing values, it has a considerable

number of outliers in the numerical variables such as BILL_AMT, PAY_AMT etc. Using decision tree is all the more right and justified because of the presence of outliers.

The decision tree model is implemented by setting up the training and validation datasets. A 70%-30% split is used for splitting the dataset into training and validation. The main intention of the validation dataset is to check for the accuracy of the model built using training data and prevent overfitting.

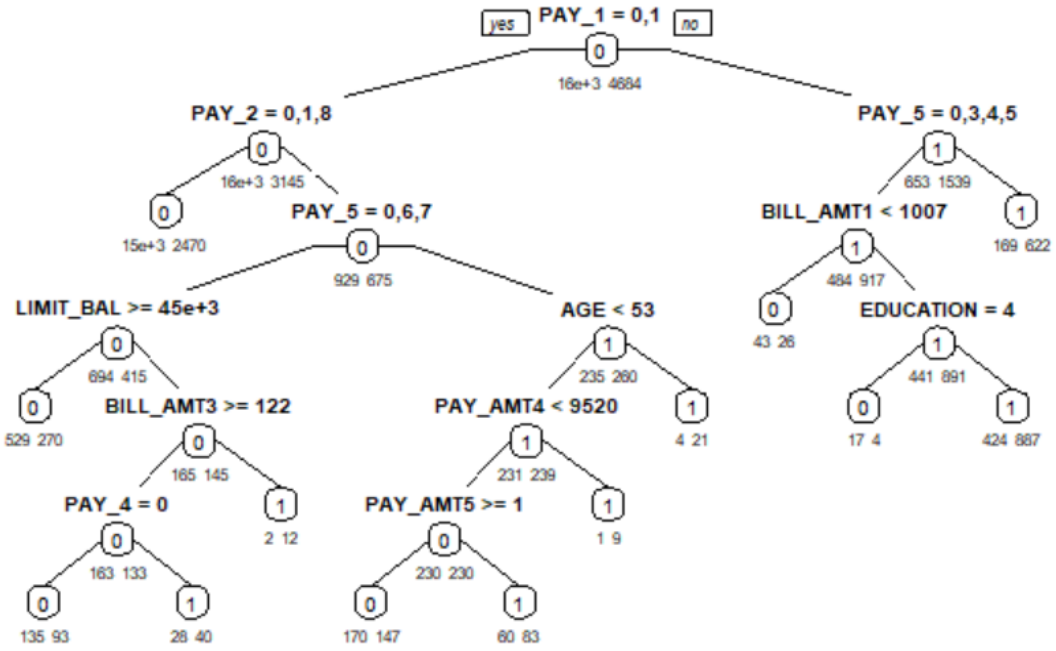
While building the decision tree model, there were 2 parameters to optimize for my dataset. They were C_p, the complexity parameter and 'maxdepth' parameter. Both the parameters influence the extent of data pruning and levels of splitting. The complexity parameter controls the decision to go for a split, meaning it will allow splitting only if the overall R^2 increases by at least C_p. The `maxdepth` parameter determines the depth of tree splits.

The C_p parameter was chosen to be 0.001 on a trial-and-error basis looking at training and validation accuracy. The values of `maxdepth` was also chosen to be 4 in the same fashion. A working snapshot of training accuracy and validation accuracy for various values of `maxdepth` is shown below as a working reference.

maxdepth	Training Accuracy	Validation Accuracy
2	0.819	0.821
3	0.821	0.822
4	0.822	0.822
5	0.822	0.822
6	0.824	0.822

Beyond the value of 4, we see that the validation accuracy decreases although training accuracy increases, and this is a case of overfitting. Hence, `maxdepth` is given an optimized value of 4.

The decision tree model that is built is visually realized using the below tree plot. At the top we have PAY_1 variable that decides the first split for the data. Any data point sequentially flows through the tree until it reaches a leaf node at the end, which eventually is classified as 0 or 1.



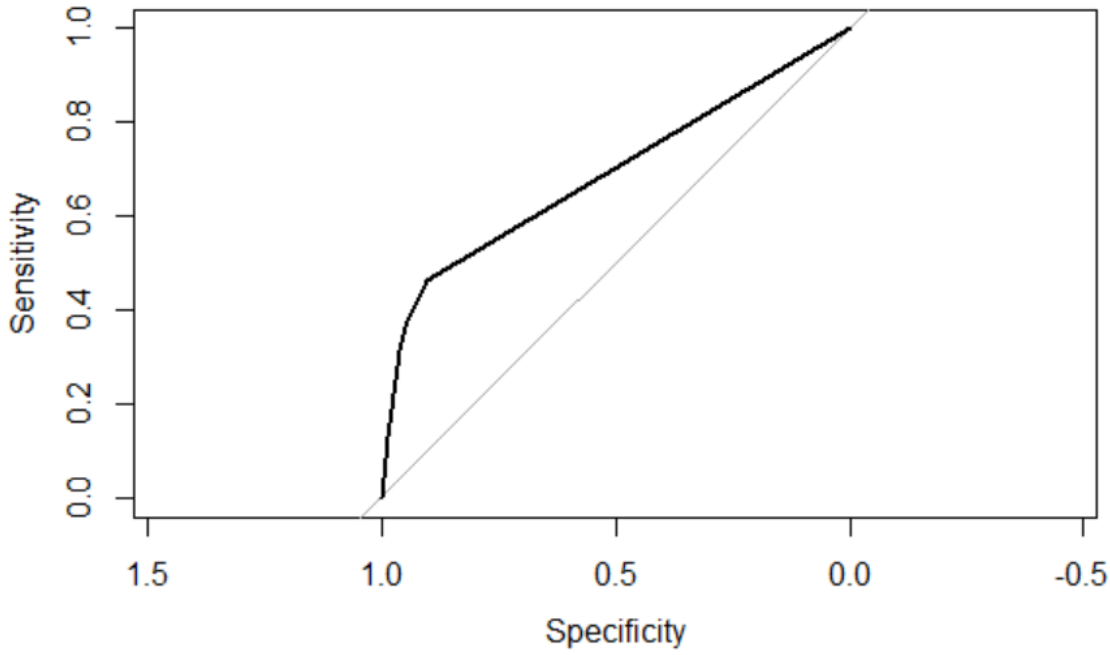
The model that is built is tested for accuracy after which, we evaluate the model's prediction accuracy over the validation dataset. The confusion matrices are computed for both the datasets. The ROC curve for the validation dataset is also plotted for evaluation purposes.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15488  2915
         1   828  1769

               Accuracy : 0.822
                 95% CI : (0.817, 0.827)
    No Information Rate : 0.777
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.389
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.949
            Specificity : 0.378
         Pos Pred Value : 0.842
         Neg Pred Value : 0.681
             Prevalence : 0.777
         Detection Rate : 0.738
   Detection Prevalence : 0.876
      Balanced Accuracy : 0.663

       'Positive' Class : 0
```

Confusion Matrix for Training data

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6686 1240
         1  362  712

               Accuracy : 0.822
                 95% CI : (0.814, 0.83)
    No Information Rate : 0.783
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.374
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.949
            Specificity : 0.365
         Pos Pred Value : 0.844
         Neg Pred Value : 0.663
             Prevalence : 0.783
         Detection Rate : 0.743
   Detection Prevalence : 0.881
      Balanced Accuracy : 0.657

       'Positive' Class : 0
```

Confusion Matrix for Validation data



Area under the curve: 0.691

From the above results, it can be inferred that we have a comparable value of test accuracy which means that the decision tree model is good enough to predict payment defaults with an accuracy of 82%. The ROC index is 0.691. The ROC index will be used an evaluation metric for comparison with other classification models.

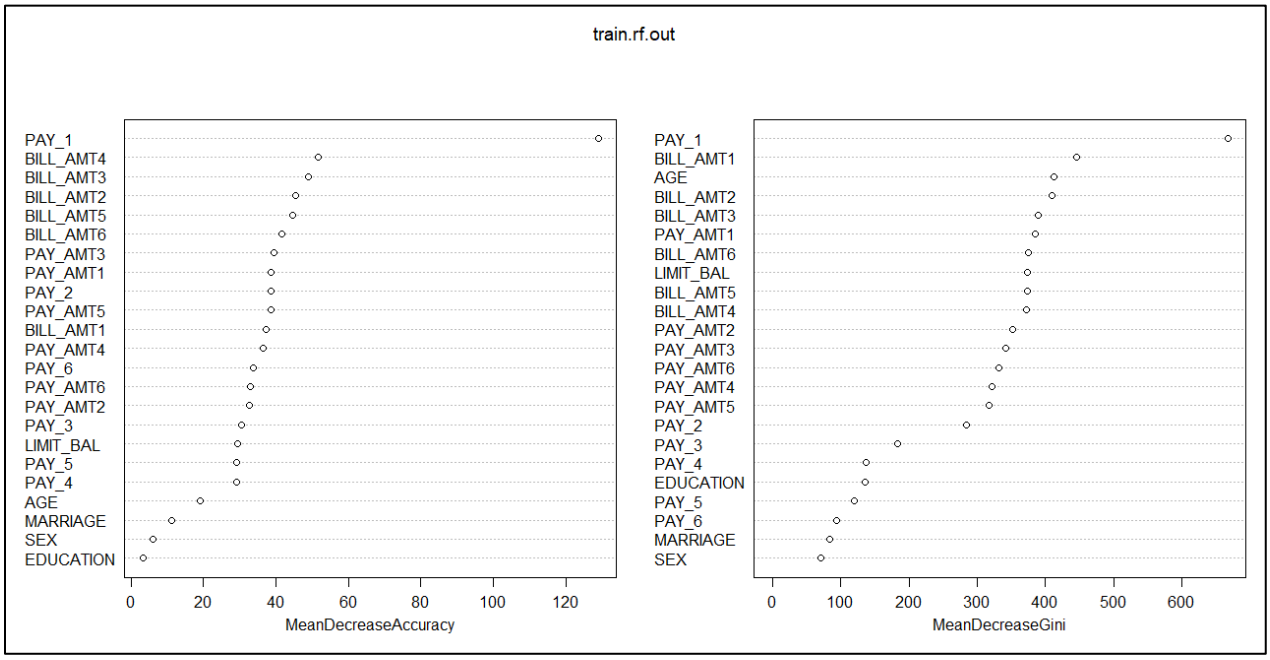
2. Random Forest

Radom forest classifier is called as an ensemble algorithm. Ensembled algorithms are those which combine several other algorithms that could be similar or different to classify data. Random forest classifier, as an ensemble algorithm, is an extension of decision tree model, in the sense that it creates a set of decision trees from a random subset of the training dataset and aggregates the votes from different decision trees to decide the final class of the test data.

Implementing the random forest classifier for the give dataset, we split 70% of the data as training dataset and 30% of data as validation dataset. All the predictor variables are used for building the model. The importance of predictor variables is plotted below which is inherent in Random Forest model implementation.

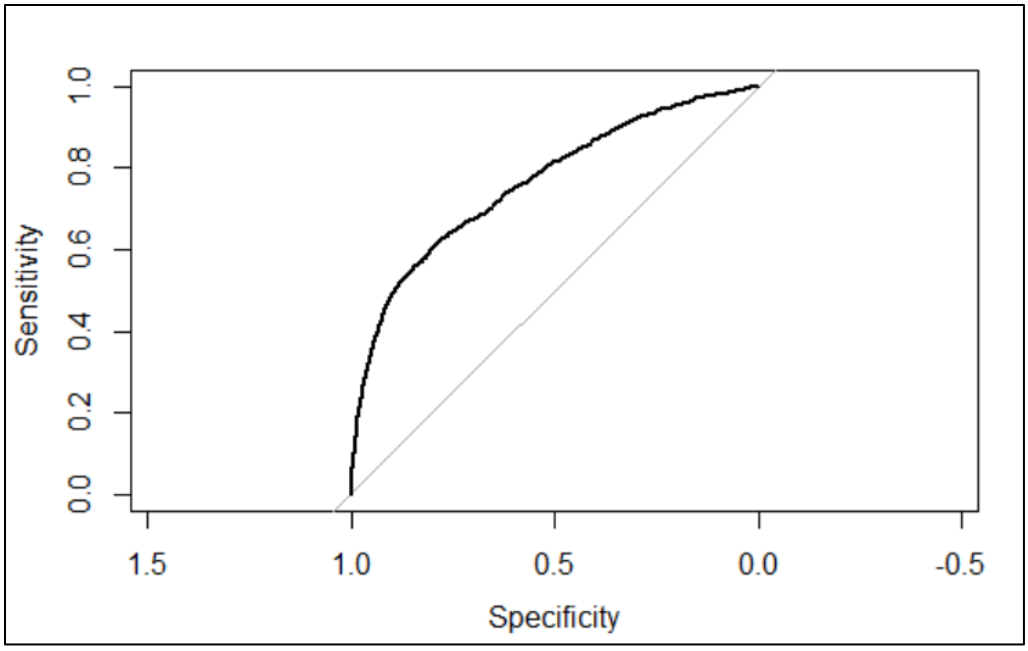
The below plot shows us that PAY_1 is the most significant predictor followed by BILL_AMT4, BILL_AMT3 and so on. The importance is plotted with Gini Index as the base,

which is a statistical measure of distribution indicating inequality. Higher the Gini, greater the inequality.



The confusion matrix for the test set and the ROC index are computed. The results are pasted below.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6591 1254
         1  401  754

               Accuracy : 0.816
                 95% CI : (0.808, 0.824)
    No Information Rate : 0.777
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.375
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.943
            Specificity : 0.375
         Pos Pred Value : 0.840
         Neg Pred Value : 0.653
             Prevalence : 0.777
         Detection Rate : 0.732
   Detection Prevalence : 0.872
      Balanced Accuracy : 0.659

       'Positive' Class : 0
```




```
Area under the curve: 0.764
```

It can be noted that the test accuracy is recorded as 81.6% from the Confusion matrix and the Area under the curve is 0.764.

3. Logistic Regression

Logistic regression extends the ideas of linear regression to the situation where the outcome variable, Y, is categorical. In this scenario, Logistic Regression is used for classification. The target variable is a binary variable and hence, there are only 2 possible outcomes, is there a default or not.

In Logistic Regression, we always check for missing values because when we lose data points in Regression, we lose power and sample size thereby introducing a possibility of a potential bias. Missing values are generally imputed appropriately whenever regression is used for classification. In this dataset, since there are no missing values, no imputation is required, and we directly go ahead with data partitioning and model building.

The same 70% - 30% split is applied to split the dataset into training and validation datasets. Then, using the training dataset, we perform logistic regression and observe the summaries. The significant predictors are looked up after which, similar to the decision tree model evaluation, the confusion matrices are computed for the training and validation datasets and the ROC index is calculated for comparison with other classification models. The results are pasted below.

```
Call:
glm(formula = default.payment.next.month ~ ., family = "binomial",
    data = train2.df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.430  -0.594  -0.525  -0.338   3.107

Coefficients: (1 not defined because of singularities)
                 Estimate     Std. Error z value             Pr(>|z|)
(Intercept)   -1.550215067    0.109823857  -14.12 < 0.0000000000000002 ***
LIMIT_BAL     -0.000001045    0.000000187   -5.59          0.000000023 ***
SEX2          -0.162472762    0.038241293   -4.25          0.000021509 ***
EDUCATION2    -0.027003487    0.043875618   -0.62              0.53825
EDUCATION3    -0.134019730    0.059797911   -2.24              0.02501 *
EDUCATION4    -1.027369505    0.231035899   -4.45          0.000008716 ***
MARRIAGE2     -0.156580406    0.043178282   -3.63              0.00029 ***
MARRIAGE3     -0.050509848    0.168135088   -0.30              0.76386
AGE            0.005673502    0.002345323    2.42              0.01556 *
PAY_11         0.816428526    0.059225456   13.79 < 0.0000000000000002 ***
PAY_12         2.059359290    0.066209984   31.10 < 0.0000000000000002 ***
PAY_13         2.041317929    0.176213267   11.58 < 0.0000000000000002 ***
PAY_14         1.914536946    0.357583073    5.35          0.000000086 ***
PAY_15         1.192728992    0.580866174    2.05              0.04004 *
PAY_16        -0.033514154    1.070607041   -0.03              0.97503
PAY_17         1.208561129    1.509089956    0.80              0.42322
PAY_18       -11.316245005  273.277547336   -0.04              0.96697
PAY_21        -0.488088413    0.568165328   -0.86              0.39031
PAY_22         0.136680014    0.071812920    1.90              0.05700 .
PAY_23         0.003031065    0.178304038    0.02              0.98644
PAY_24        -0.657251860    0.364893529   -1.80              0.07167 .
PAY_25         1.857511820    0.862324745    2.15              0.03123 *
PAY_26         1.806176904    1.588718630    1.14              0.25559
PAY_27                  NA             NA      NA                  NA
PAY_28        13.130163773  639.927506571    0.02              0.98363
PAY_31       -12.107503865  535.411476706   -0.02              0.98196
PAY_32         0.309957681    0.071209605    4.35          0.000013444 ***
PAY_33         0.520689760    0.230873774    2.26              0.02411 *
PAY_34        -0.359769206    0.495860174   -0.73              0.46812
PAY_35        -0.925468176    1.012822328   -0.91              0.36085
PAY_36        13.940841157  273.278055524    0.05              0.95931
PAY_37         0.387934680    0.876229243    0.44              0.65796
PAY_38        24.198133034  607.906700343    0.04              0.96825
PAY_41        27.316962499  757.185873336    0.04              0.97122
PAY_42         0.256972368    0.079069251    3.25              0.00115 **
PAY_43         0.153497139    0.266807927    0.58              0.56508
PAY_44         0.118613726    0.519671151    0.23              0.81945
PAY_45        -1.205911177    0.888321034   -1.36              0.17462
PAY_46       -27.835875465  350.489034325   -0.08              0.93670
PAY_47       -35.168933894  462.529482626   -0.08              0.93939
PAY_48       -61.714198093  729.653540495   -0.08              0.93260
PAY_52         0.250086154    0.086021603    2.91              0.00365 **
```

```
PAY_53          0.029829089    0.251451015    0.12             0.90557
PAY_54         -0.148602535    0.545664137   -0.27             0.78537
PAY_55          1.513806023    1.016474572    1.49             0.13642
PAY_56         24.531483434  362.005198438    0.07             0.94597
PAY_57         46.341023115  495.713403622    0.09             0.92552
PAY_58         54.389594850 1345.778636396    0.04             0.96776
PAY_62          0.319508685    0.075315960    4.24         0.000022130 ***
PAY_63          0.719665492    0.256902566    2.80             0.00509 **
PAY_64          0.164723926    0.642929392    0.26             0.79779
PAY_65         -0.537268657    0.876839539   -0.61             0.54005
PAY_66          1.179413046    1.013048267    1.16             0.24433
PAY_67        -10.727994972  178.314442899   -0.06             0.95203
PAY_68         -4.189143032  747.607199597   -0.01             0.99553
BILL_AMT1      -0.000002554    0.000001268   -2.01             0.04400 *
BILL_AMT2       0.000003494    0.000001690    2.07             0.03870 *
BILL_AMT3      -0.000000188    0.000001588   -0.12             0.90554
BILL_AMT4       0.000000626    0.000001591    0.39             0.69386
BILL_AMT5       0.000001420    0.000001741    0.82             0.41476
BILL_AMT6      -0.000002570    0.000001381   -1.86             0.06276 .
PAY_AMT1       -0.000012003    0.000002632   -4.56         0.000005117 ***
PAY_AMT2       -0.000005445    0.000002300   -2.37             0.01793 *
PAY_AMT3       -0.000003997    0.000002071   -1.93             0.05358 .
PAY_AMT4       -0.000006663    0.000002460   -2.71             0.00675 **
PAY_AMT5       -0.000001329    0.000002018   -0.66             0.51013
PAY_AMT6       -0.000003261    0.000001664   -1.96             0.05010 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22201  on 20999  degrees of freedom
Residual deviance: 18482  on 20934  degrees of freedom
AIC: 18614

Number of Fisher Scoring iterations: 12
```

Looking at the summary output of the Logistic Regression model, it can be inferred that there are several significant predictors for the target variable. Regression in Rstudio converts categorical variables into dummy classes for performing and analysis. Some of the significant predictors include, as listed above, include LIMIT_BAL, SEX2, EDUCATION4 and MARRIAGE2. The SEX2 denotes the Female class, EDUCATION4 denotes the other educational levels and MARRIAGE2 denotes Singles. It is a good time to recall that some of the significant predictors such as SEX2, EDUCATION4 etc. that are listed here were already predicted as significantly influential variables while EDA was performed.

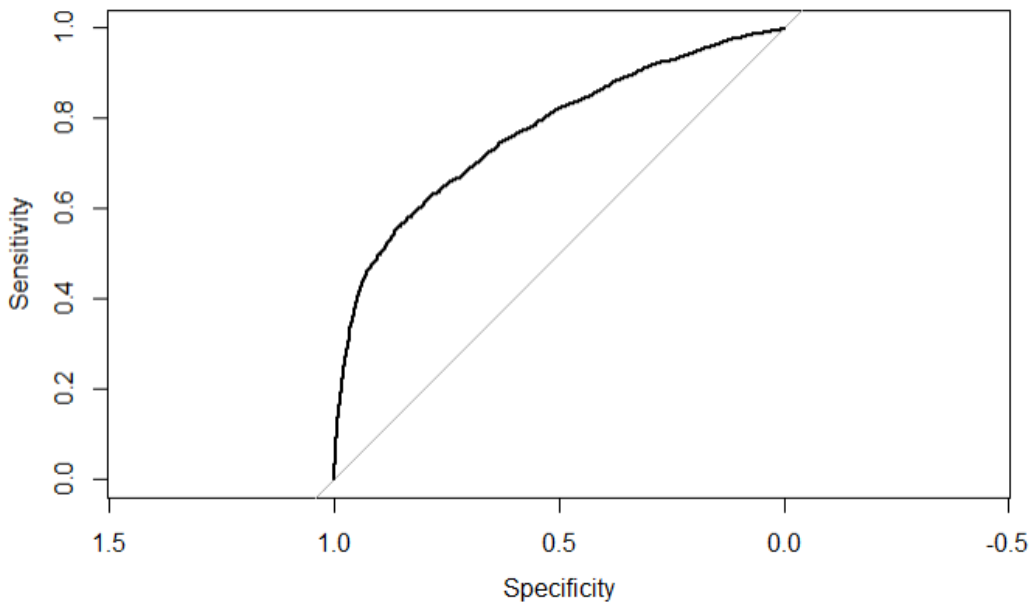
Using confusion matrix, the accuracy of the trained model is determined as 81.1% and when we test the trained model on the validation dataset, the validation accuracy is determined to be 81.7%. The validation accuracy is higher than the training accuracy in this scenario. The confusion matrices are pasted below.

Confusion Matrix for Training data

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15909  3526
         1   443  1122

               Accuracy : 0.811
                 95% CI : (0.806, 0.816)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.281
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.973
            Specificity : 0.241
         Pos Pred Value : 0.819
         Neg Pred Value : 0.717
             Prevalence : 0.779
         Detection Rate : 0.758
   Detection Prevalence : 0.925
      Balanced Accuracy : 0.607

       'Positive' Class : 0
```

Confusion matrix for Validation data

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6839 1474
         1  173  514

               Accuracy : 0.817
                 95% CI : (0.809, 0.825)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.306
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.975
            Specificity : 0.259
         Pos Pred Value : 0.823
         Neg Pred Value : 0.748
             Prevalence : 0.779
         Detection Rate : 0.760
   Detection Prevalence : 0.924
      Balanced Accuracy : 0.617

       'Positive' Class : 0
```




```
Area under the curve: 0.769
```

The ROC curve is plotted, and the AUC value is obtained to be 0.769, which is higher than the AUC obtained for Decision tree and Random Forest models.

4. Support Vector Machine

Support Vector Machine classification or SVM classification (in short) is termed as a black-box classification method because the process of transformation of the input to the output is obfuscated from the user. SVMs are primarily employed for binary classification. They try to separate the data points with one class on one side and the other class on the other side, making them as far as possible.

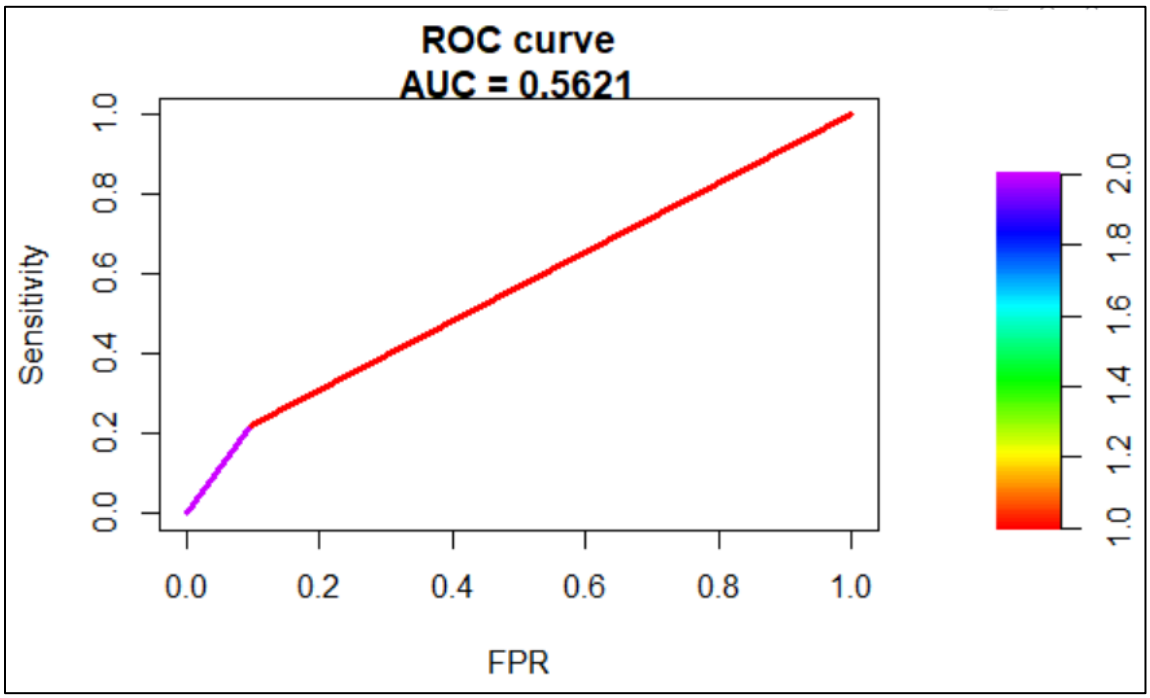
In this project, SVM is employed for the same reason, to perform binary classification on the credit card dataset. Like the above model implementations, the data is split into training (70%) and validation (30%) datasets. Since, the dataset is dense with almost 23 predictor variables, unlike other model implementations, where all predictor variables were considered, we only consider a subset of the predictor variables for training the SVM model. The Radial Basis Kernel function is used to transform the linearly not separable data. The summary output, confusion matrix for the validation data and the ROC index are calculated. The results are pasted below.

```
Call:
svm(formula = default.payment.next.month ~ LIMIT_BAL + SEX +
    EDUCATION + MARRIAGE + AGE + PAY_1 + PAY_2 + PAY_3 +
    PAY_4 + PAY_5, data = svm_train.df, kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.02083

Number of Support Vectors:  8464

 ( 4507 3957 )


Number of Classes:  2

Levels:
 0 1
```

The SVM model reports that the it is a C-type classification and the kernel used is Radial as mentioned already. Importantly, the hyperparameter values are listed, namely, Cost = 1, Gamma = 0.02083. Cost hyperparameter controls the extent of allowed misclassification while Gamma controls the extent of influence of the Kernel function on the data points. Higher values of Cost parameter indicate lower misclassification and narrower decision boundaries, while high values of Gamma indicate that decision boundaries would be more wiggled.

```
Confusion Matrix and Statistics


pred1    0    1
    0 6756 1371
    1  253  619

               Accuracy : 0.82
                 95% CI : (0.811, 0.827)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.344
 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.964
            Specificity : 0.311
         Pos Pred Value : 0.831
         Neg Pred Value : 0.710
             Prevalence : 0.779
         Detection Rate : 0.751
   Detection Prevalence : 0.903
      Balanced Accuracy : 0.637

       'Positive' Class : 0
```

The confusion matrix indicates a comparable accuracy of 82% on the validation dataset. The ROC curve plot and the AUC value is given below. AUC is recorded as 0.5621 which is quite low compared to a base value of 0.5.



Conclusion

Four different classification models have been built namely Decision tree model, Random forest model, Logistic Regression and Support Vector Machine. The results and the performance metrics are captured for the same. A quick summary of the results is pasted below for reference.

Model	Test Accuracy (%)	ROC index / AUC
Decision Tree	82.2	0.691
Random Forest	81.6	0.764
Logistic Regression	81.7	0.769
Support Vector Machine	82.0	0.562

The ROC index is chosen to be the evaluation metric for assessing the performance of the built models overall. The ROC index is considered as the evaluation metric because considering the test accuracy is highly misleading at times, especially when the class of interest is occurring only in meagre amounts. Though the class of interest here, which is that a customer would default, occurs relatively higher equaling 22.12%, it is better to consider the ROC index as the evaluation metric because it also gives a good sense of how the model is performing from Sensitivity and Specificity perspectives i.e. how good the model is in identifying positives as positives and negatives as negatives.

Looking at the summary above, it can be concluded that Logistic Regression is the best model out of all the others considering the ROC index. Though Logistic Regression has a lower test accuracy, its higher ROC index signifies that the model has learnt better and can distinguish better when compared to other models. A perfect illustration of misleading accuracy is the SVM model here which has a good test accuracy but a poor ROC index.

The built Logistic Regression model can be used as the basis for prediction of payment defaults among credit card customers. The model can also be used to serve bigger populations and provide predictions with good accuracy and minimum error. The model's predictions can be looked upon as the basis for improvement of financial measures and eligibility standards that could help avoid such a financial debt crisis again. Ensembled models can be built to study diverse populations and factors thereby helping serve the society better with analytics and classification techniques.

References

Taiwan's Credit Card Crisis. (2019). Retrieved July 15, 2019 from https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/

Jian Sun. *Analyzing Default Payments of Credit Card Clients in Taiwan.* (2019). Retrieved July 11, 2019 from https://www.researchgate.net/publication/311714926_Analyzing_Default_Payments_of_Credit_Card_Clients_in_Taiwan

Tae Soo Kang, Guonan Ma. *Recent episodes of credit card distress in Asia*. (2017). Retrieved July 11, 2019 from https://pdfs.semanticscholar.org/ff79/93d0ae5017cd074d3939554bbd2dbd24dc70.pdf

Theos Evgeniou, Spyros Zoumpoulis. *Classification for Credit Card Default.* (2019). Retrieved June 24, 2019 from http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/ClassificationProcessCreditCardDefault.html#step_4:_classification_and_interpretation

Savan Patel. *Random Forest Classifier*. (2019). Retrieved July 8, 2019 from https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1