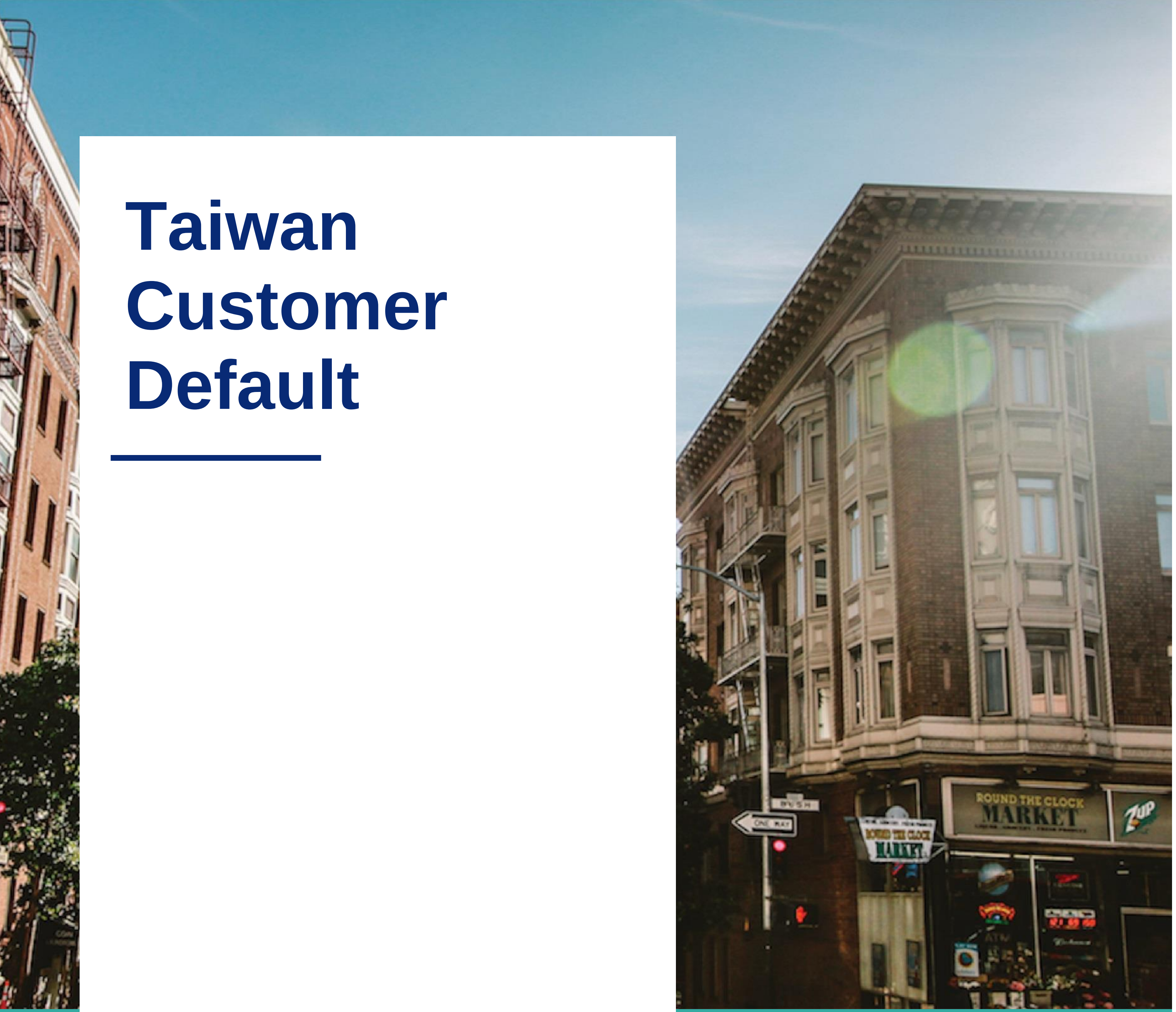
Taiwan Customer Default



MARCH 14, 2020

Authored by: Syed Shoaib Mohammad

Introduction

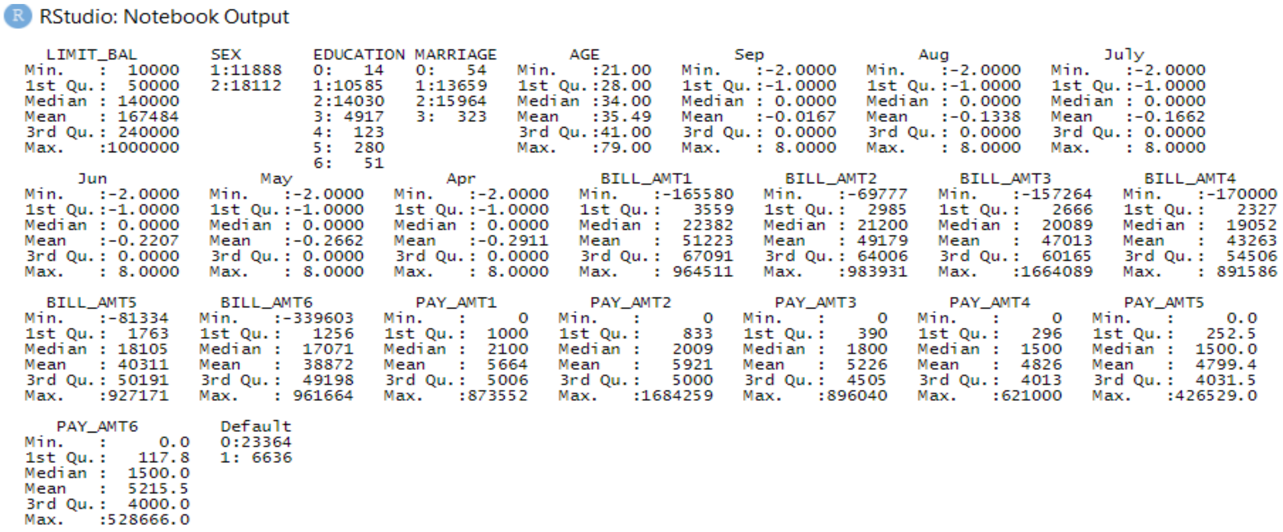
In 2005, Taiwan was confronted with a cash card crisis. Through the year of 2005, the total default volume in Taiwan amounted to over 30 billion NT$. The number of over-borrowing people was about 300,000 to 400,000(almost 1.7% of the total population and 2% of adult population). The average loan value was more than 1.9 million NT$. The outbreak of debt crisis exerted an overwhelming influence on Taiwan's economic development and social stability.

Currently I work as a part of lending dept. for a reputed international bank, I have always been intrigued by the decision making abilities of the bank which were tailor made as a reference while discuss a lending product with a prospective customer, I am very much aware of the various factors involved to judge a credibility of a particular applicant, but was not aware of the math or the science behind the credit scoring, internal scoring, the question asked and the answers given during initial stages of an application. I also deal with existing customer who have defaulted and asses their financial circumstance to set up a plan to retrieve the defaulted amount. The need of this project is to gain a deeper understanding of the analytics behind judging the credibility of a particular person based on the past behavior in terms of repayments, and bill amounts and amount repaid for the respective bills and other information's held by the bank or a third-party agency.

"In this case study we would try to explore what were the traits of these over-borrowers and also come up with an internal scoring mechanism and recommendations that could have been used for controlling bad debts".

EDA

The data consists of 30000 observations and 24 variables (no missing values), will be considering the default as the dependent variable and the remaining as the predictor variables. Some variables such as sex, marriage, education and Default have been converted to factor variable for the ease of analysis and history of past repayments have been renamed with the name of that respective month as a reference. A quick view of the 5-point summary.

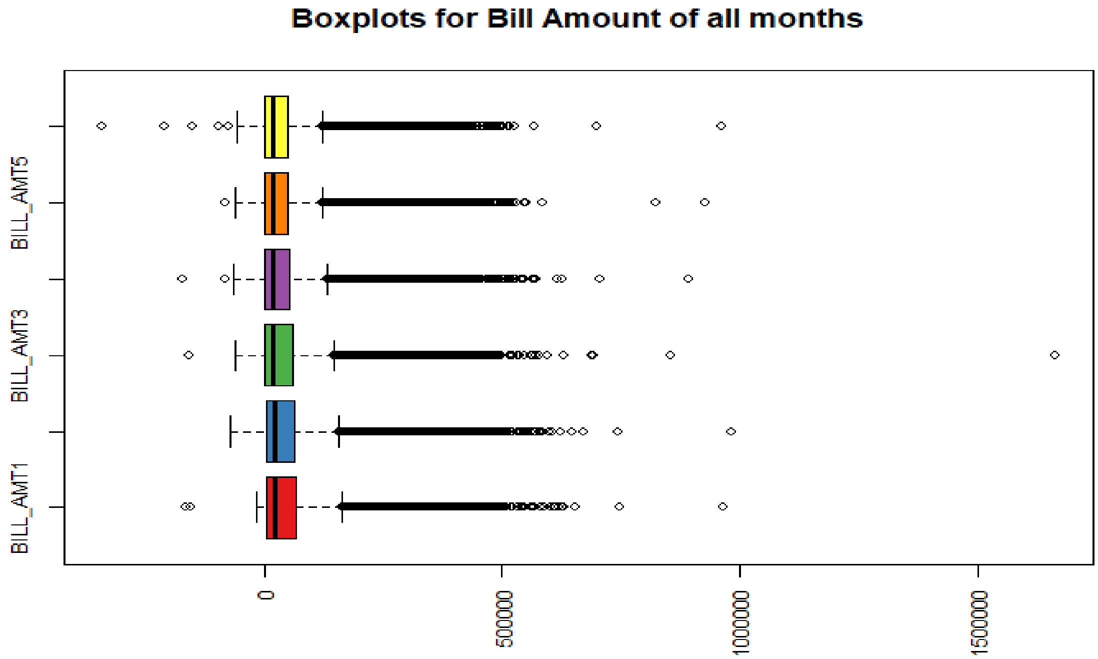


```
   LIMIT_BAL        SEX        EDUCATION MARRIAGE      AGE                 Sep                   Aug                  July         
 Min.   :  10000   1:11888   0:   14    0:   54   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000  
 1st Qu.:  50000   2:18112   1:10585    1:13659   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000  
 Median : 140000             2:14030    2:15964   Median :34.00   Median : 0.0000   Median : 0.0000   Median : 0.0000  
 Mean   : 167484             3: 4917    3:  323   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338   Mean   :-0.1662  
 3rd Qu.: 240000             4:  123              3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
 Max.   :1000000             5:  280              Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000  
                             6:   51                                                                                   
      Jun                 May                  Apr              BILL_AMT1          BILL_AMT2          BILL_AMT3           BILL_AMT4      
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000  
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327  
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :  22382   Median : 21200   Median :  20089   Median :  19052  
 Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911   Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263  
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506  
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586  

   BILL_AMT5         BILL_AMT6           PAY_AMT1           PAY_AMT2           PAY_AMT3          PAY_AMT4          PAY_AMT5       
 Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0   Min.   :     0.0  
 1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5  
 Median : 18105   Median :  17071   Median :  2100   Median :   2009   Median :  1800   Median :  1500   Median :  1500.0  
 Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826   Mean   :  4799.4  
 3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5  
 Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000   Max.   :426529.0  

    PAY_AMT6        Default 
 Min.   :     0.0   0:23364  
 1st Qu.:   117.8   1: 6636  
 Median :  1500.0            
 Mean   :  5215.5            
 3rd Qu.:  4000.0            
 Max.   :528666.0            
```

Continuous Variables Analysis:

- The variable Limit Balance is skewed to the positive side.
- The variable Bill amount 1, Bill amount 2, Bill amount 3, Bill amount 4, Bill amount 5, Bill amount 6, have majority of observations between -100000 to 200000 (we see a negative bill amount may be due to overspending against the available limit).
- Box Plots show that there are outliers in all the Bill Amounts for all months the median seems to be even across all months and the IQR range also seems narrow.

- Box Plots for Payment amount has extremely narrow IQR.
- Overall default rate in the data is 22%

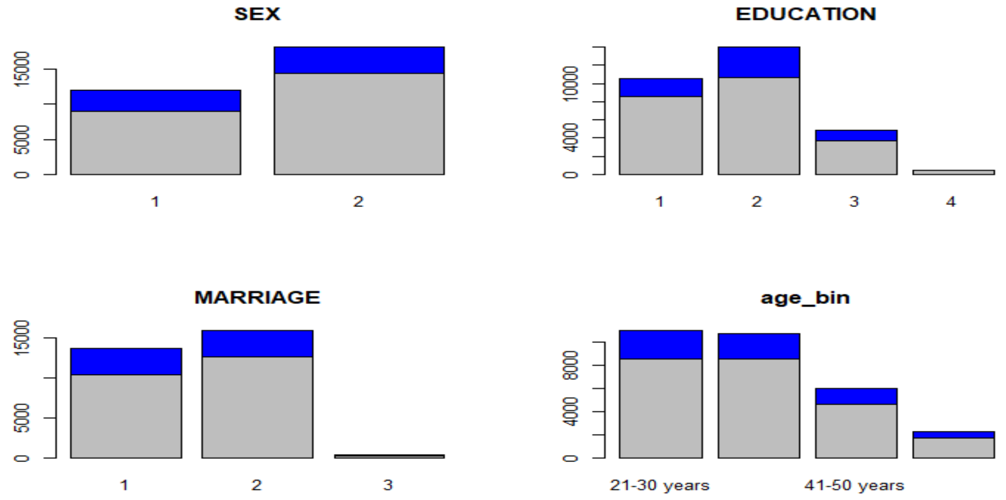


Categorical Variable Analysis:

Factor variables like Education and Marriage need to be collapsed to reduce the levels. No signs of any missing values. The past repayments have been collapsed to 6 levels, marriage to 3 levels, education to 4 levels and an additional variable age_bin can be created for further analysis.

- In terms of education University and High school have a higher contribution towards default with 23% and 25% respectively in the given category.
- Males have higher default rate of nearly 24% in the given category.
- Marital status as Married and others have a higher percentage of default in the given category with 23% approx., but observation for status others is less.

- Default rate contribution for each age bucket is between 20-25%, highest is for age bucket 51-79 years, and average Limit Balance is high for age bucket 31-40 years.

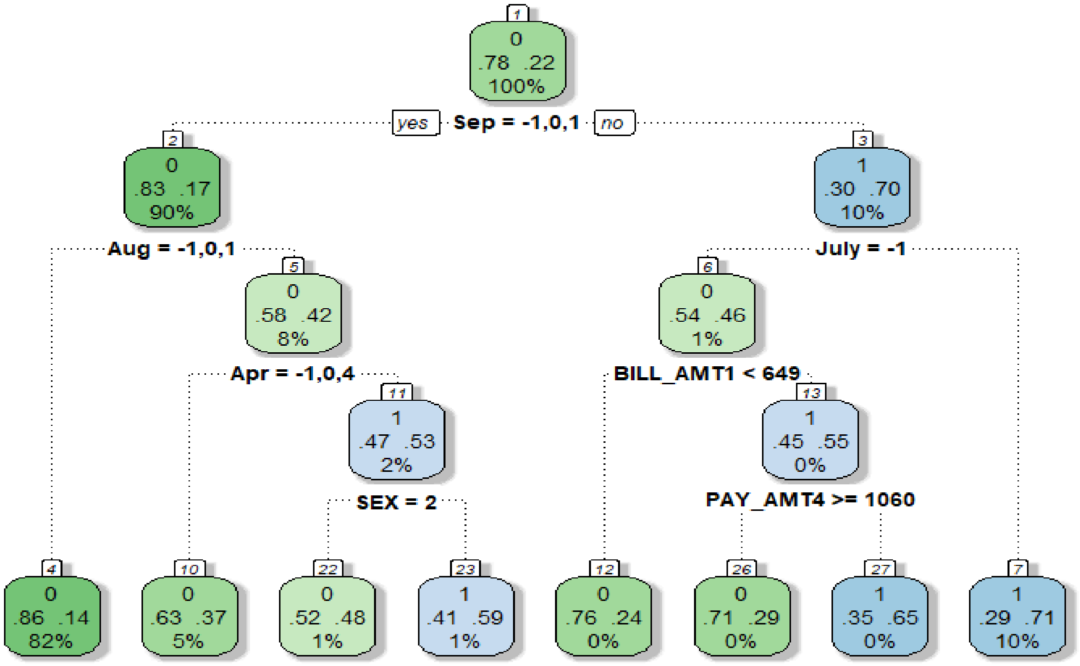

SEX

EDUCATION

MARRIAGE

age_bin

Decision tree to understand the information gain based on variable:



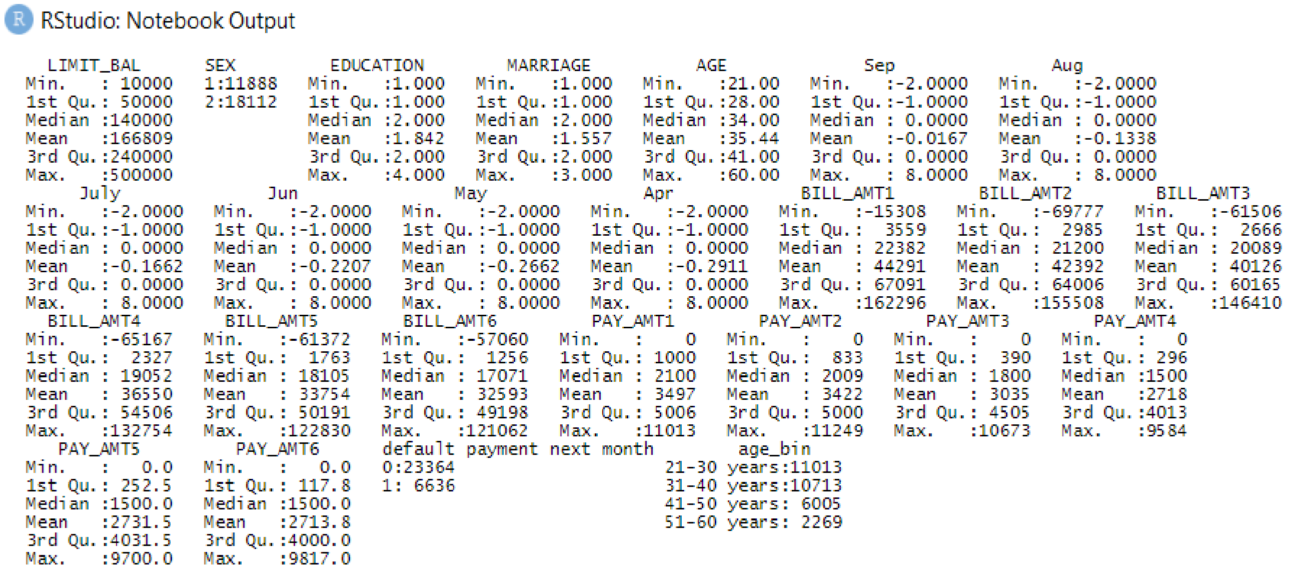
Rattle 2020-Jan-12 03:20:13 Shoaib

Interpretation:

- A customer who has been late in making repayments in the past at least by 2 months as it stands in the month of September along with customers repayment status who have not payed duly in the month of July and those who did pay, if their billed amount is greater than 649NT September and have made a payment greater than or equal to 1060 NT$ the month of July, tend to default on their payments.
- Customer who have payed duly along with those who have not been late (more than a month) as it stands in the month of September but have been late in repayments as it stands in August and April by at least 2 months especially Males, tend to default on their payments.

Data Pre-Processing

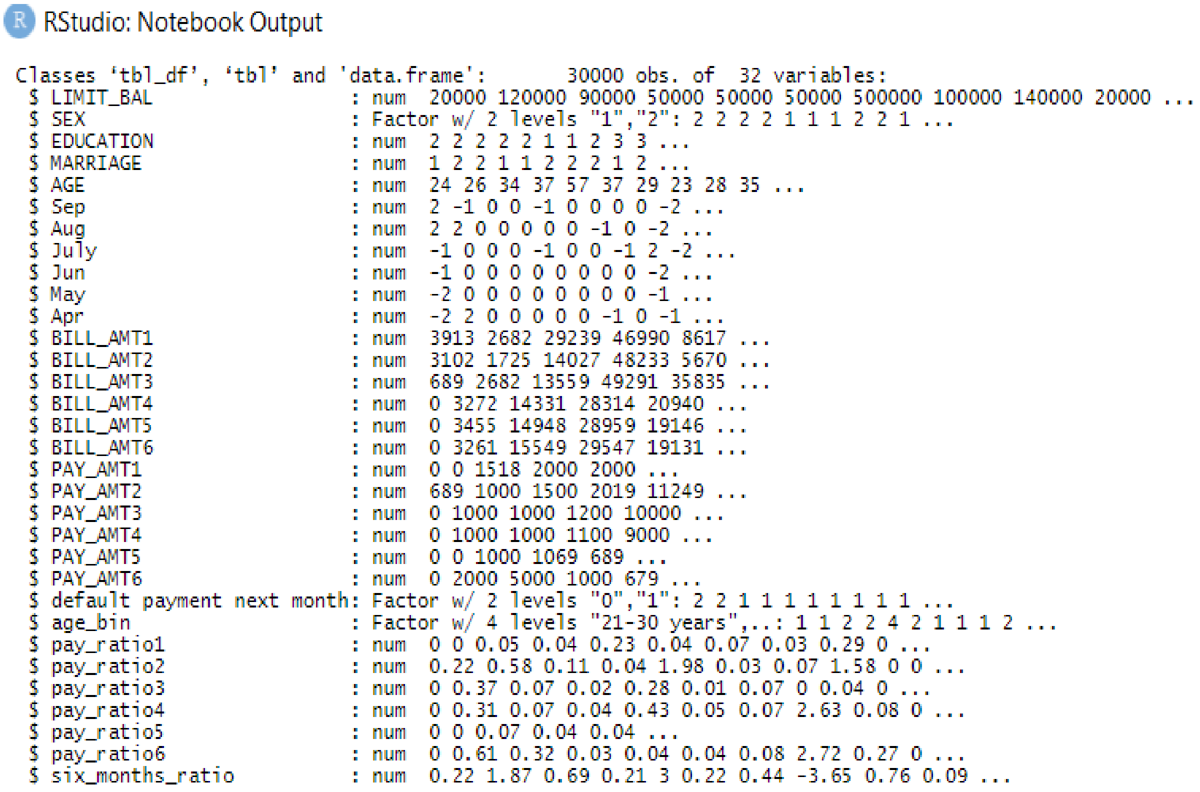
- At first, we shall go ahead and clean the data by eliminating outliers using the capping technique by defining the outlier boundaries, where anything below 0.25-1.5IQR and 0.75+1.5IQR will termed as an outlier and will be replaced by the min and max of the respective boundaries. Below is the summary of the data after having treated the outliers.



```
   LIMIT_BAL        SEX          EDUCATION         MARRIAGE          AGE             Sep               Aug         
 Min.   : 10000   1:11888   Min.   :1.000   Min.   :1.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000  
 1st Qu.: 50000   2:18112   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000  
 Median :140000             Median :2.000   Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000  
 Mean   :166809             Mean   :1.842   Mean   :1.557   Mean   :35.44   Mean   :-0.0167   Mean   :-0.1338  
 3rd Qu.:240000             3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
 Max.   :500000             Max.   :4.000   Max.   :3.000   Max.   :60.00   Max.   : 8.0000   Max.   : 8.0000  
      July              Jun                May                Apr             BILL_AMT1        BILL_AMT2        BILL_AMT3     
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-15308   Min.   :-69777   Min.   :-61506  
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:  3559   1st Qu.:  2985   1st Qu.:  2666  
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 22382   Median : 21200   Median : 20089  
 Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911   Mean   : 44291   Mean   : 42392   Mean   : 40126  
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 67091   3rd Qu.: 64006   3rd Qu.: 60165  
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   :162296   Max.   :155508   Max.   :146410  
   BILL_AMT4        BILL_AMT5        BILL_AMT6         PAY_AMT1         PAY_AMT2         PAY_AMT3         PAY_AMT4   
 Min.   :-65167   Min.   :-61372   Min.   :-57060   Min.   :    0   Min.   :    0   Min.   :    0   Min.   :   0  
 1st Qu.:  2327   1st Qu.:  1763   1st Qu.:  1256   1st Qu.: 1000   1st Qu.:  833   1st Qu.:  390   1st Qu.: 296  
 Median : 19052   Median : 18105   Median : 17071   Median : 2100   Median : 2009   Median : 1800   Median :1500  
 Mean   : 36550   Mean   : 33754   Mean   : 32593   Mean   : 3497   Mean   : 3422   Mean   : 3035   Mean   :2718  
 3rd Qu.: 54506   3rd Qu.: 50191   3rd Qu.: 49198   3rd Qu.: 5006   3rd Qu.: 5000   3rd Qu.: 4505   3rd Qu.:4013  
 Max.   :132754   Max.   :122830   Max.   :121062   Max.   :11013   Max.   :11249   Max.   :10673   Max.   :9584  
    PAY_AMT5         PAY_AMT6      default payment next month           age_bin     
 Min.   :   0.0   Min.   :   0.0   0:23364                       21-30 years:11013  
 1st Qu.: 252.5   1st Qu.: 117.8   1: 6636                       31-40 years:10713  
 Median :1500.0   Median :1500.0                                 41-50 years: 6005  
 Mean   :2731.5   Mean   :2713.8                                 51-60 years: 2269  
 3rd Qu.:4031.5   3rd Qu.:4000.0                                                    
 Max.   :9700.0   Max.   :9817.0                                                    
```

Creating new variables in terms of ratios:

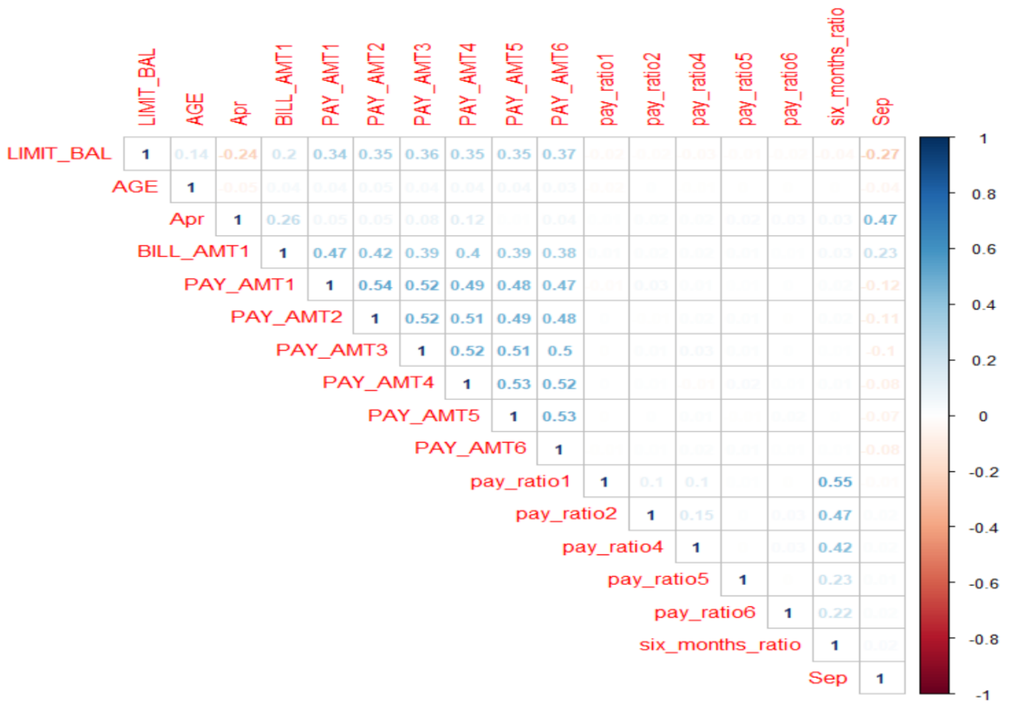
New variables payment ratio has been added onto the data set by taking the ratio of the payment amount to the bill amount of the respective months and a sum of all ratio named as six months ratio, Not sure if this would help as the interest rates have not been provided in the data set.



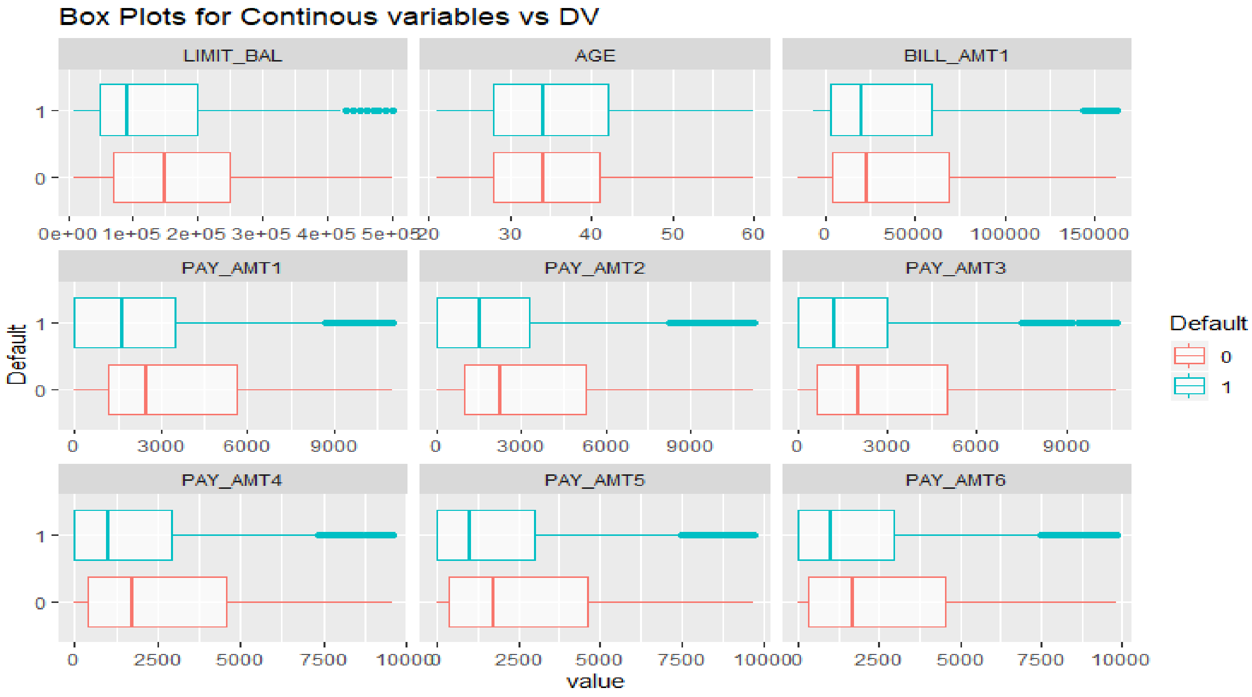
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	30000 obs. of  32 variables:
 $ LIMIT_BAL                 : num  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
 $ SEX                       : Factor w/ 2 levels "1","2": 2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : num  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : num  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : num  24 26 34 37 57 37 29 23 28 35 ...
 $ Sep                       : num  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ Aug                       : num  2 2 0 0 0 0 0 -1 0 -2 ...
 $ July                      : num  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ Jun                       : num  -1 0 0 0 0 0 0 0 0 -2 ...
 $ May                       : num  -2 0 0 0 0 0 0 0 0 -1 ...
 $ Apr                       : num  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : num  3913 2682 29239 46990 8617 ...
 $ BILL_AMT2                 : num  3102 1725 14027 48233 5670 ...
 $ BILL_AMT3                 : num  689 2682 13559 49291 35835 ...
 $ BILL_AMT4                 : num  0 3272 14331 28314 20940 ...
 $ BILL_AMT5                 : num  0 3455 14948 28959 19146 ...
 $ BILL_AMT6                 : num  0 3261 15549 29547 19131 ...
 $ PAY_AMT1                  : num  0 0 1518 2000 2000 ...
 $ PAY_AMT2                  : num  689 1000 1500 2019 11249 ...
 $ PAY_AMT3                  : num  0 1000 1000 1200 10000 ...
 $ PAY_AMT4                  : num  0 1000 1000 1100 9000 ...
 $ PAY_AMT5                  : num  0 0 1000 1069 689 ...
 $ PAY_AMT6                  : num  0 2000 5000 1000 679 ...
 $ default payment next month: Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 ...
 $ age_bin                   : Factor w/ 4 levels "21-30 years",..: 1 1 2 2 4 2 1 1 1 2 ...
 $ pay_ratio1                : num  0 0 0.05 0.04 0.23 0.04 0.07 0.03 0.29 0 ...
 $ pay_ratio2                : num  0.22 0.58 0.11 0.04 1.98 0.03 0.07 1.58 0 0 ...
 $ pay_ratio3                : num  0 0.37 0.07 0.02 0.28 0.01 0.07 0 0.04 0 ...
 $ pay_ratio4                : num  0 0.31 0.07 0.04 0.43 0.05 0.07 2.63 0.08 0 ...
 $ pay_ratio5                : num  0 0 0.07 0.04 0.04 ...
 $ pay_ratio6                : num  0 0.61 0.32 0.03 0.04 0.04 0.08 2.72 0.27 0 ...
 $ six_months_ratio          : num  0.22 1.87 0.69 0.21 3 0.22 0.44 -3.65 0.76 0.09 ...
```

Eliminating Multicollinearity:

Based on correlation plot, will be eliminating Multicollinearity keeping in mind the important variables from the decision tree. Below are the final plots.

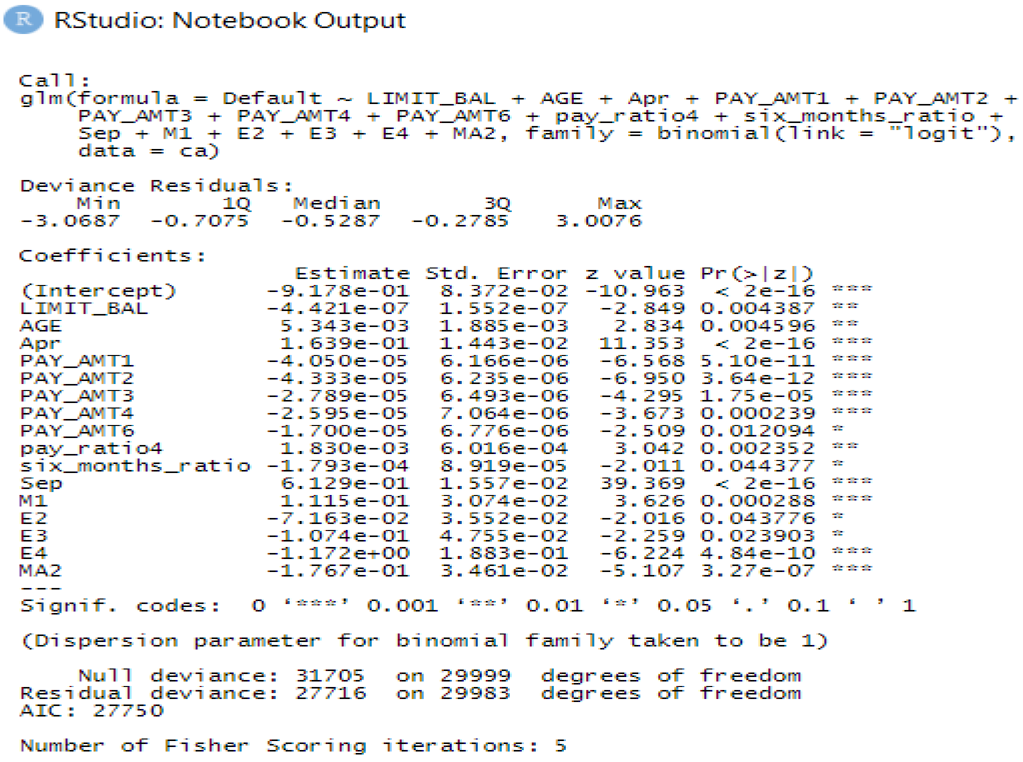


Boxplot on the remaining variables against Default (Excluding the ratios):



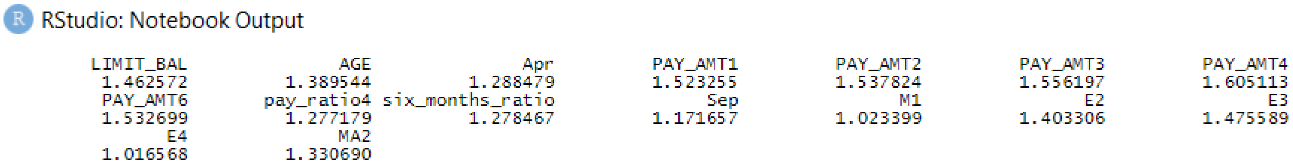
The plot above shows that the IQR range for Defaulters is lesser as compared to those who are non-defaulters along with the median value except for the variable “AGE” which seems to be the same for both defaulters and non-defaulters.

Creating dummy variable for the factor variables and building a logistic regression Model and VIF check:



RStudio: Notebook Output

```
Call:
glm(formula = Default ~ LIMIT_BAL + AGE + Apr + PAY_AMT1 + PAY_AMT2 +
    PAY_AMT3 + PAY_AMT4 + PAY_AMT6 + pay_ratio4 + six_months_ratio +
    Sep + M1 + E2 + E3 + E4 + MA2, family = binomial(link = "logit"),
    data = ca)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0687  -0.7075  -0.5287  -0.2785   3.0076

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -9.178e-01  8.372e-02 -10.963  < 2e-16 ***
LIMIT_BAL        -4.421e-07  1.552e-07  -2.849 0.004387 **
AGE               5.343e-03  1.885e-03   2.834 0.004596 **
Apr               1.639e-01  1.443e-02  11.353  < 2e-16 ***
PAY_AMT1         -4.050e-05  6.166e-06  -6.568 5.10e-11 ***
PAY_AMT2         -4.333e-05  6.235e-06  -6.950 3.64e-12 ***
PAY_AMT3         -2.789e-05  6.493e-06  -4.295 1.75e-05 ***
PAY_AMT4         -2.595e-05  7.064e-06  -3.673 0.000239 ***
PAY_AMT6         -1.700e-05  6.776e-06  -2.509 0.012094 *
pay_ratio4        1.830e-03  6.016e-04   3.042 0.002352 **
six_months_ratio -1.793e-04  8.919e-05  -2.011 0.044377 *
Sep               6.129e-01  1.557e-02  39.369  < 2e-16 ***
M1                1.115e-01  3.074e-02   3.626 0.000288 ***
E2               -7.163e-02  3.552e-02  -2.016 0.043776 *
E3               -1.074e-01  4.755e-02  -2.259 0.023903 *
E4               -1.172e+00  1.883e-01  -6.224 4.84e-10 ***
MA2              -1.767e-01  3.461e-02  -5.107 3.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31705  on 29999  degrees of freedom
Residual deviance: 27716  on 29983  degrees of freedom
AIC: 27750

Number of Fisher Scoring iterations: 5
```



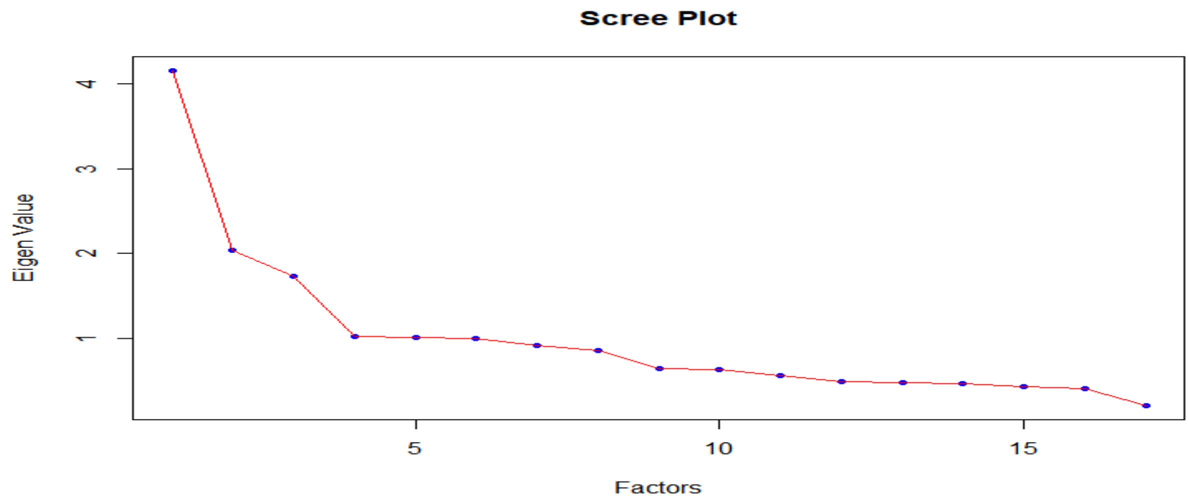
RStudio: Notebook Output

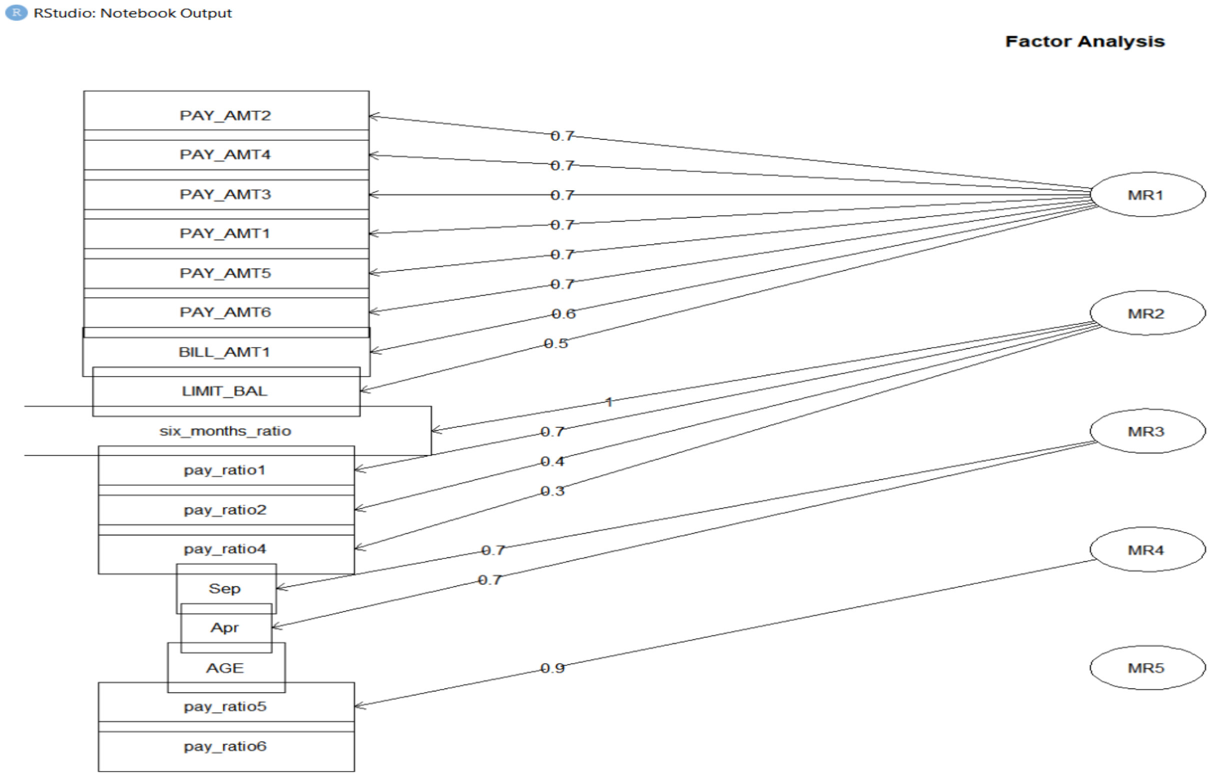
```
      LIMIT_BAL              AGE              Apr         PAY_AMT1         PAY_AMT2         PAY_AMT3         PAY_AMT4
       1.462572         1.389544         1.288479         1.523255         1.537824         1.556197         1.605113
       PAY_AMT6       pay_ratio4 six_months_ratio              Sep               M1               E2               E3
       1.532699         1.277179         1.278467         1.171657         1.023399         1.403306         1.475589
             E4              MA2
       1.016568         1.330690
```

Dimensionality Reduction:

The data was split into 70:30 ratio and PCA was done on the train and test separately using a common Mean and Standard deviation.

- The scree plot (Kaiser rule) suggests that 5 factors will be able to explain the maximum variation among attributes, considering the factors as a linear combination of the variable.





- Factor MR1 shows the correlation between attributes Credit Limit (LIMIT_BAL), Bill Amount for the month of September (BILL_AMT1), Amount payed towards the credit since April to September (PAY_AMT6-PAY_AMT1). Factor MRI can be named as behaviour/usage/score of the credit available and its usage including the repayments made towards it. We shall keep the remaining attributes as they are and bind the scores of MR1 to the train and test data for further analysis.

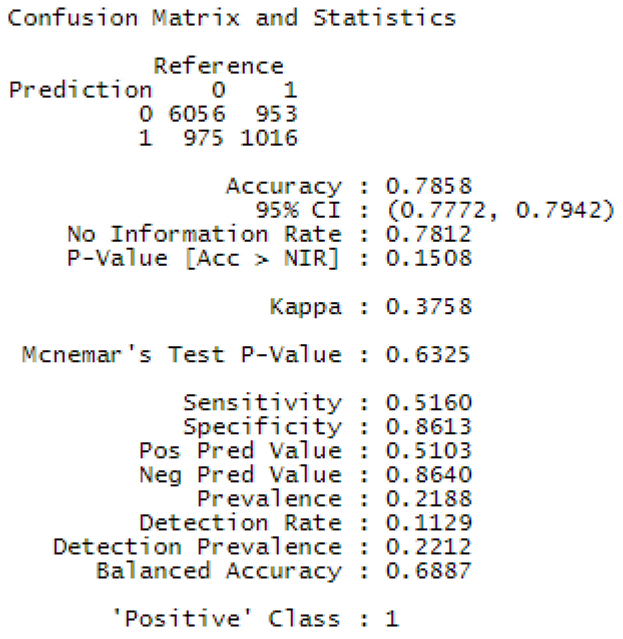
Model Building

- All the models have been built using 10-fold cross validation to reduce the chances of model overfitting. The probabilities from the training models have been fit onto the test data and have also computed an optimum threshold to most importantly increase True Positive and reduce False Negatives by computing ROC, the parameters used here to judge the best performing models are, AUC, Confusion Matrix, Gini co-efficient, sensitivity and explainability. Models used for the study are;

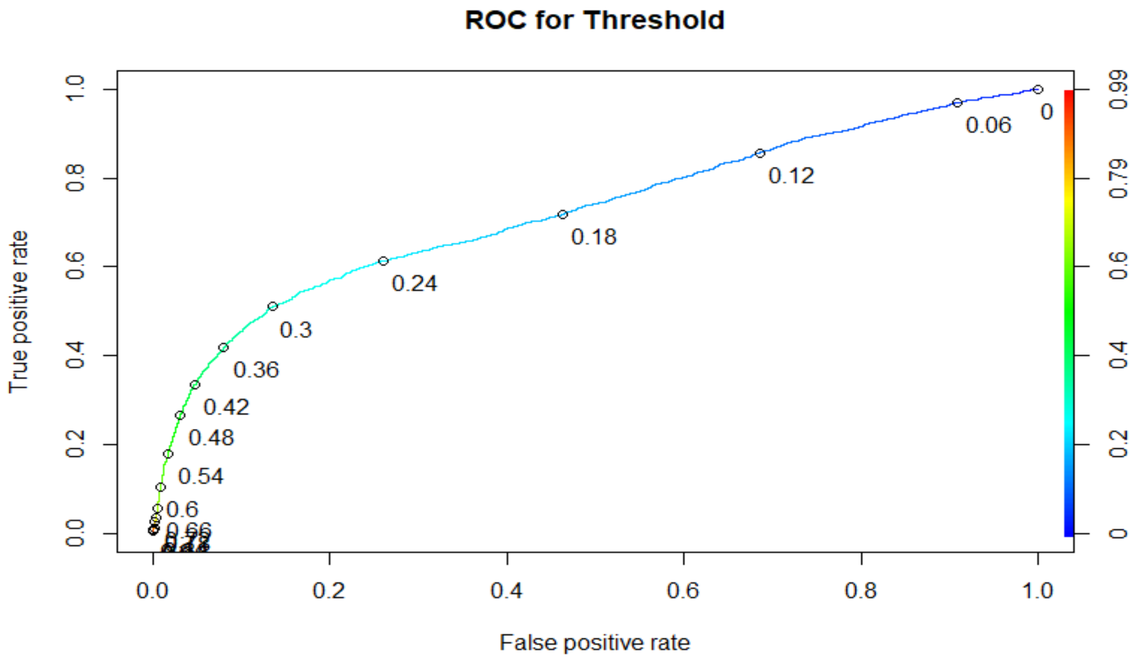
  - ***Logistic regression***
  - ***KNN***
  - ***Random Forest***
  - ***Bagging***
  - ***Boosting***
  - ***Neural Net***

Logistic Regression:

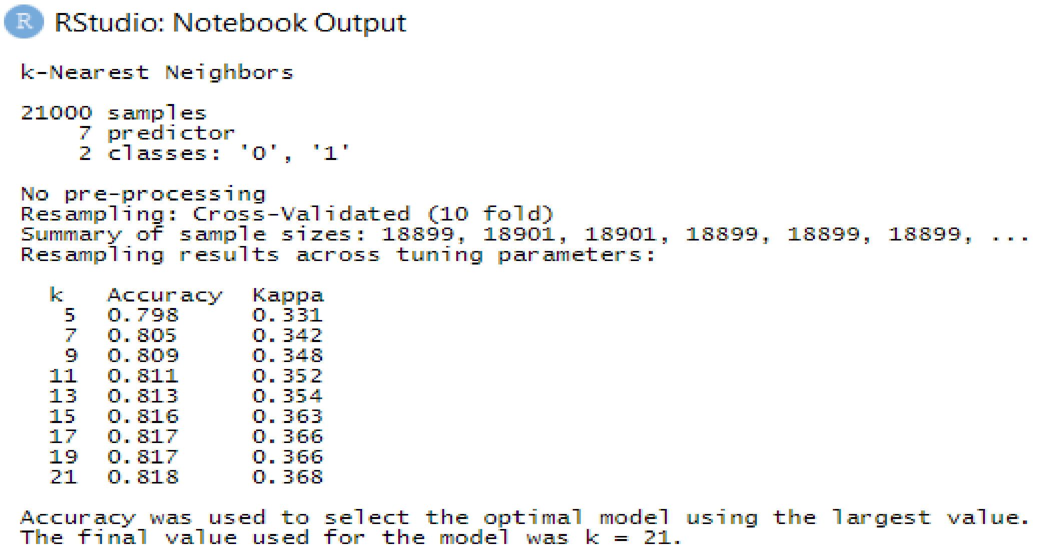
RStudio: Notebook Output



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6056  953
         1  975 1016
                                          
               Accuracy : 0.7858          
                 95% CI : (0.7772, 0.7942)
    No Information Rate : 0.7812          
    P-Value [Acc > NIR] : 0.1508          
                                          
                  Kappa : 0.3758          
                                          
 Mcnemar's Test P-Value : 0.6325          
                                          
            Sensitivity : 0.5160          
            Specificity : 0.8613          
         Pos Pred Value : 0.5103          
         Neg Pred Value : 0.8640          
             Prevalence : 0.2188          
         Detection Rate : 0.1129          
   Detection Prevalence : 0.2212          
      Balanced Accuracy : 0.6887          
                                          
       'Positive' Class : 1               
```

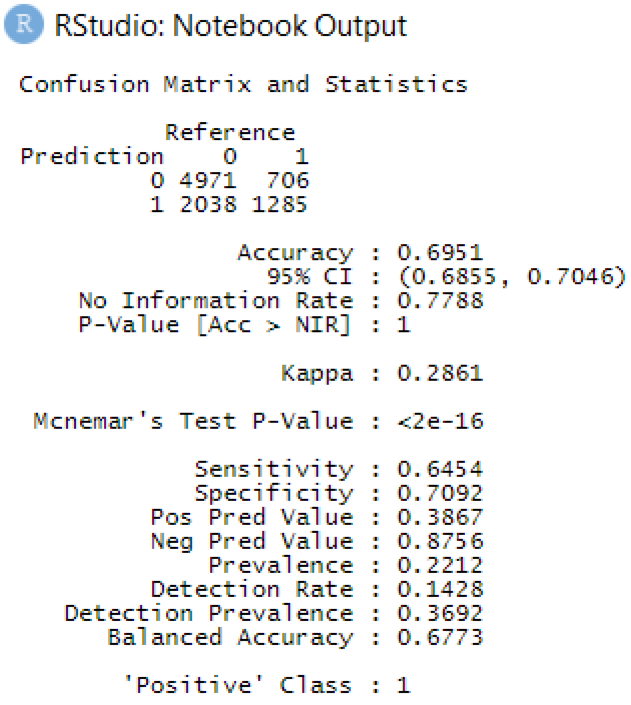


KNN:



RStudio: Notebook Output

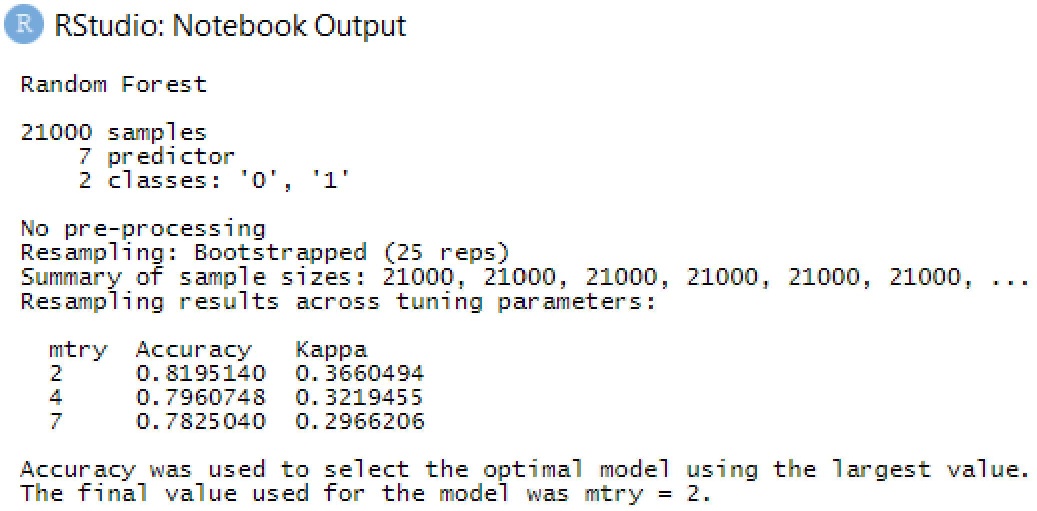
```
k-Nearest Neighbors

21000 samples
    7 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 18899, 18901, 18901, 18899, 18899, 18899, ...
Resampling results across tuning parameters:

  k   Accuracy  Kappa
   5  0.798     0.331
   7  0.805     0.342
   9  0.809     0.348
  11  0.811     0.352
  13  0.813     0.354
  15  0.816     0.363
  17  0.817     0.366
  19  0.817     0.366
  21  0.818     0.368

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 21.
```



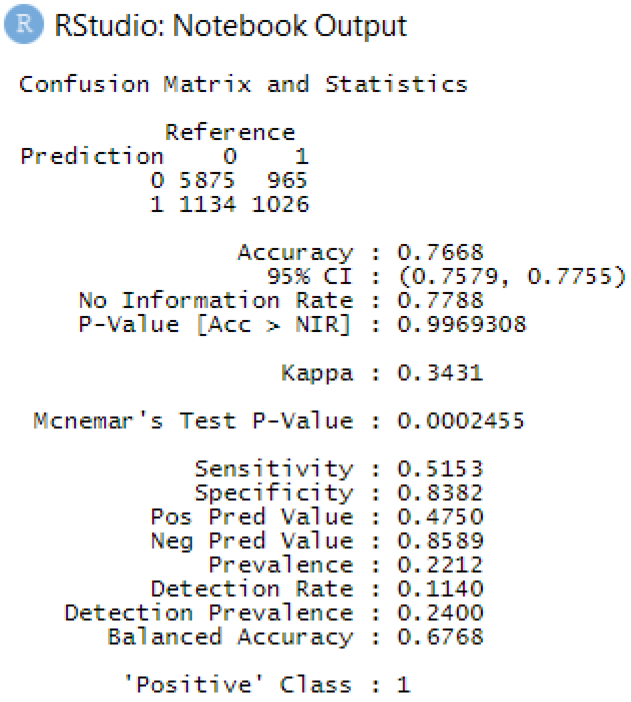
RStudio: Notebook Output

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4971  706
         1 2038 1285

               Accuracy : 0.6951
                 95% CI : (0.6855, 0.7046)
    No Information Rate : 0.7788
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2861

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6454
            Specificity : 0.7092
         Pos Pred Value : 0.3867
         Neg Pred Value : 0.8756
             Prevalence : 0.2212
         Detection Rate : 0.1428
   Detection Prevalence : 0.3692
      Balanced Accuracy : 0.6773

       'Positive' Class : 1
```

Random Forest:



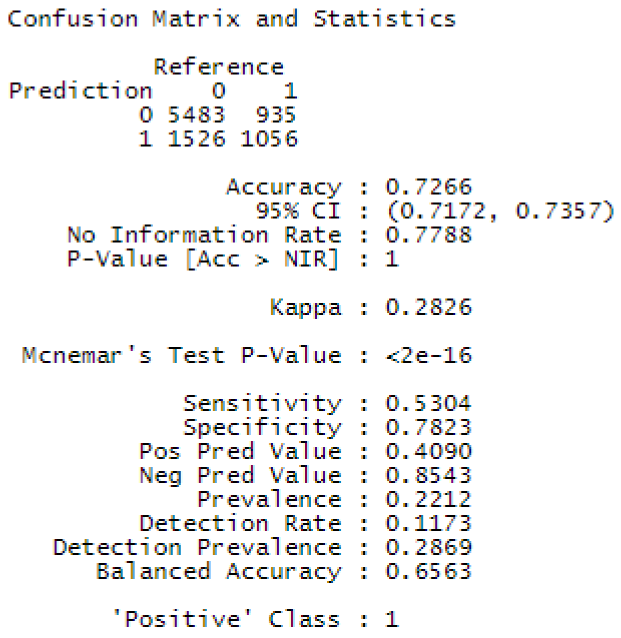
```
Random Forest

21000 samples
    7 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 21000, 21000, 21000, 21000, 21000, 21000, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  2     0.8195140  0.3660494
  4     0.7960748  0.3219455
  7     0.7825040  0.2966206

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```



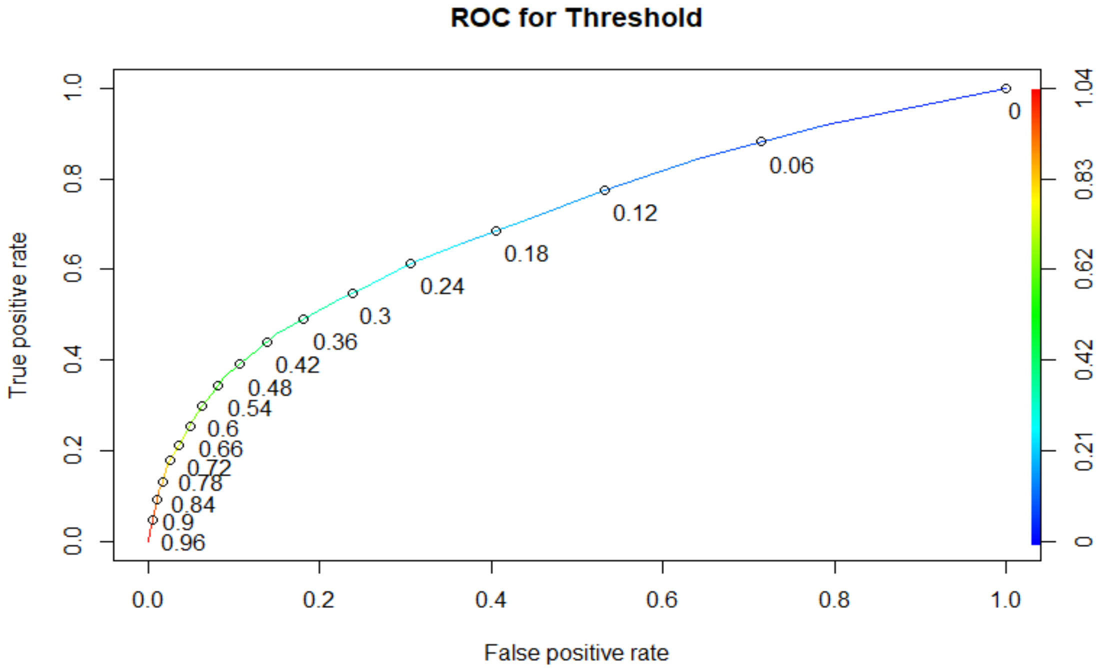
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5875  965
         1 1134 1026

               Accuracy : 0.7668
                 95% CI : (0.7579, 0.7755)
    No Information Rate : 0.7788
    P-Value [Acc > NIR] : 0.9969308

                  Kappa : 0.3431

 Mcnemar's Test P-Value : 0.0002455

            Sensitivity : 0.5153
            Specificity : 0.8382
         Pos Pred Value : 0.4750
         Neg Pred Value : 0.8589
             Prevalence : 0.2212
         Detection Rate : 0.1140
   Detection Prevalence : 0.2400
      Balanced Accuracy : 0.6768

       'Positive' Class : 1
```

Bagging:

RStudio: Notebook Output

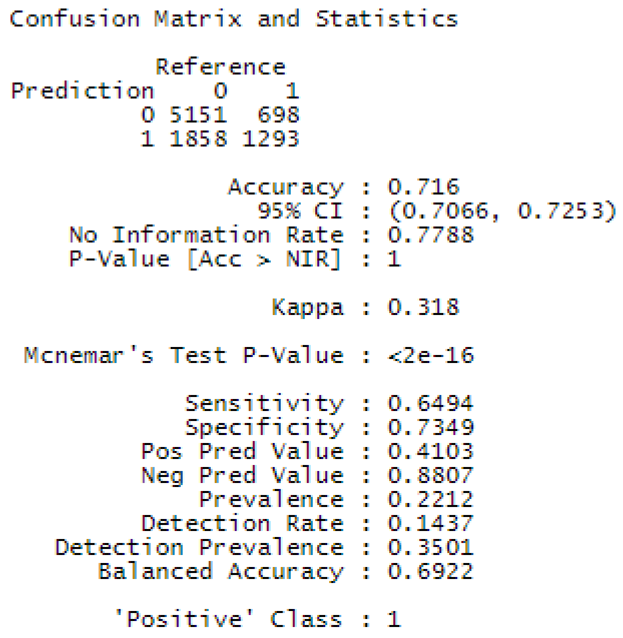


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5483  935
         1 1526 1056

               Accuracy : 0.7266
                 95% CI : (0.7172, 0.7357)
    No Information Rate : 0.7788
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2826

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5304
            Specificity : 0.7823
         Pos Pred Value : 0.4090
         Neg Pred Value : 0.8543
             Prevalence : 0.2212
         Detection Rate : 0.1173
   Detection Prevalence : 0.2869
      Balanced Accuracy : 0.6563

       'Positive' Class : 1
```

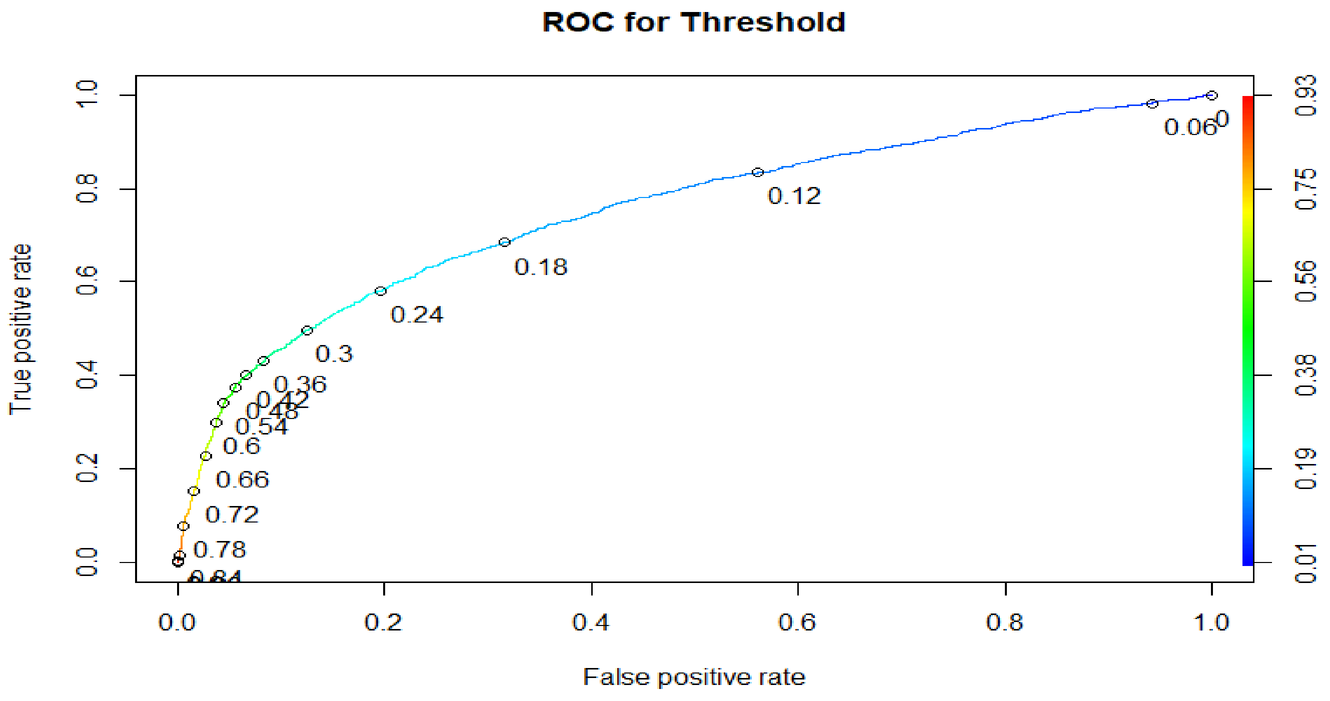


Boosting:

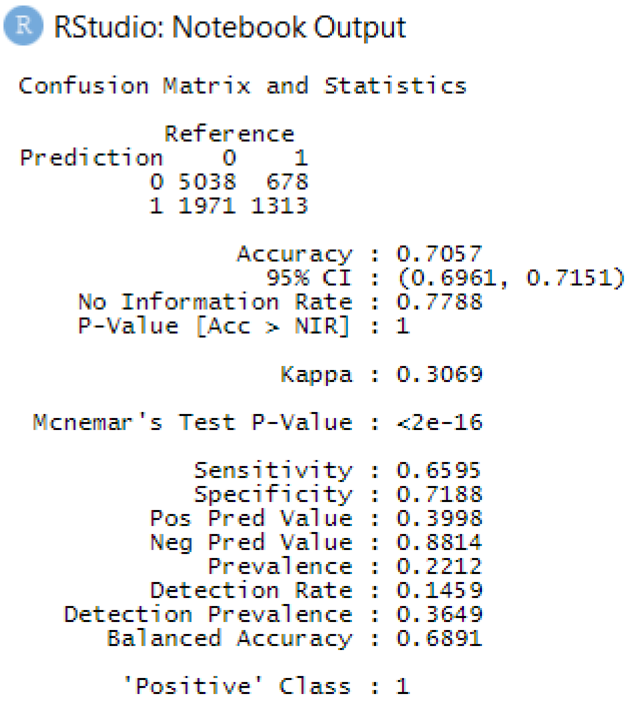
RStudio: Notebook Output



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5151  698
         1 1858 1293
                                          
               Accuracy : 0.716           
                 95% CI : (0.7066, 0.7253)
    No Information Rate : 0.7788          
    P-Value [Acc > NIR] : 1               
                                          
                  Kappa : 0.318           
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.6494          
            Specificity : 0.7349          
         Pos Pred Value : 0.4103          
         Neg Pred Value : 0.8807          
             Prevalence : 0.2212          
         Detection Rate : 0.1437          
   Detection Prevalence : 0.3501          
      Balanced Accuracy : 0.6922          
                                          
       'Positive' Class : 1               
```

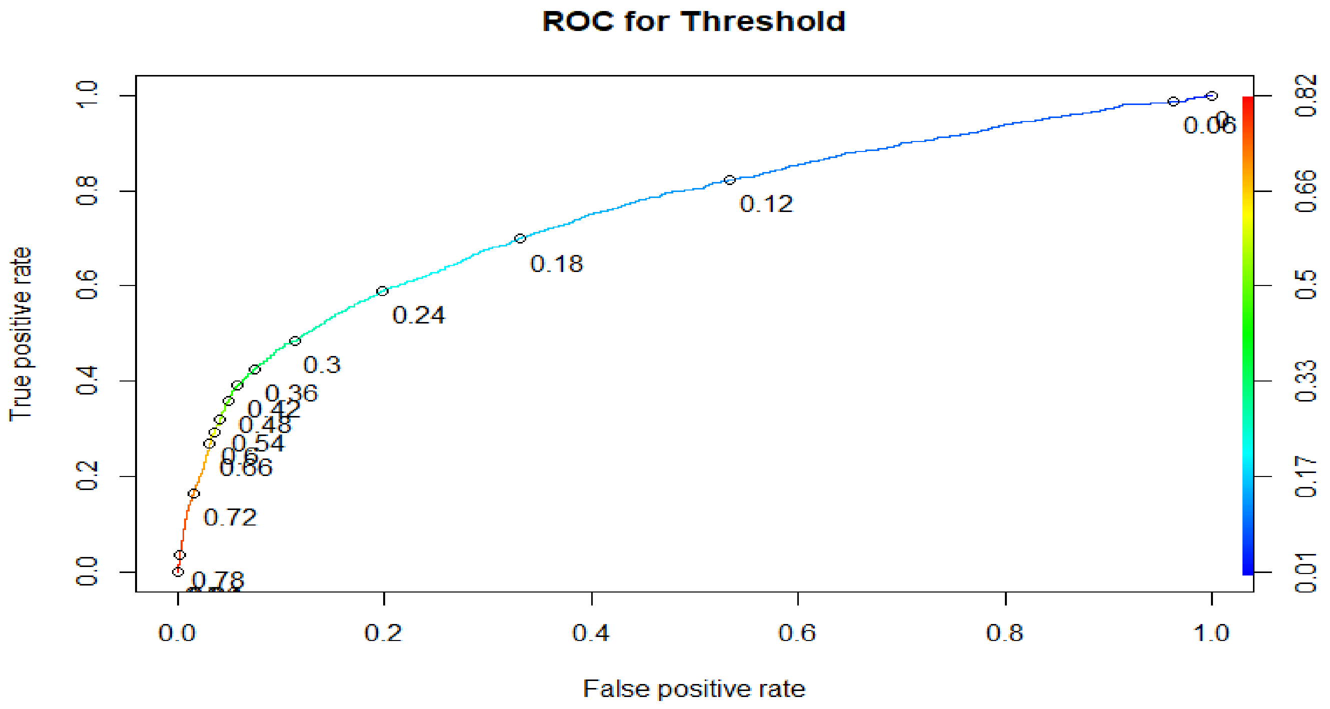


Neural Net:



RStudio: Notebook Output

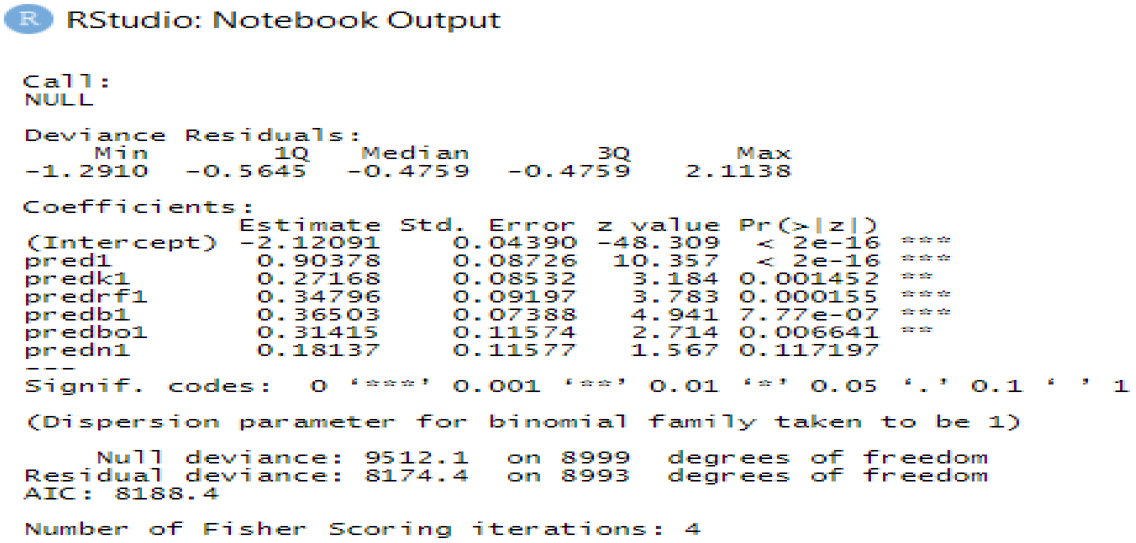
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5038  678
         1 1971 1313

               Accuracy : 0.7057
                 95% CI : (0.6961, 0.7151)
    No Information Rate : 0.7788
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3069

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6595
            Specificity : 0.7188
         Pos Pred Value : 0.3998
         Neg Pred Value : 0.8814
             Prevalence : 0.2212
         Detection Rate : 0.1459
   Detection Prevalence : 0.3649
      Balanced Accuracy : 0.6891

       'Positive' Class : 1
```



Model Comparisons:

Model	Confusion Matrix	AUC	Gini-Coefficient	Variable Importance
Logistic Regression	Accuracy:78% Sensitivity:51% Threshold:30%	72%	44%	-Sep -Credit Score
KNN	Accuracy:69% Sensitivity:64% Threshold:20%	73%	47%	-Sep -Credit Score
Random Forest	Accuracy:76% Sensitivity:51% Threshold:10%	74%	48%	-Sep -Credit Score
Bagging	Accuracy:72% Sensitivity:53% Threshold:30%	70%	41%	-Credit Score -Age
Boosting	Accuracy:71% Sensitivity:64% Threshold:20%	75%	50%	-Sep -Credit Score
Neural Net	Accuracy:70% Sensitivity:65% Threshold:20%	75%	50%	-Sep -Education as “others”

Model Blending using logistic Regression:



```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2910  -0.5645  -0.4759  -0.4759   2.1138

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12091    0.04390 -48.309  < 2e-16 ***
pred1        0.90378    0.08726  10.357  < 2e-16 ***
predk1       0.27168    0.08532   3.184 0.001452 **
predrf1      0.34796    0.09197   3.783 0.000155 ***
predb1       0.36503    0.07388   4.941 7.77e-07 ***
predbo1      0.31415    0.11574   2.714 0.006641 **
predn1       0.18137    0.11577   1.567 0.117197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9512.1  on 8999  degrees of freedom
Residual deviance: 8174.4  on 8993  degrees of freedom
AIC: 8188.4

Number of Fisher Scoring iterations: 4
```

Model Performance:

In terms of sensitivity Neural Networks is the best with a True Positive Rate of 65%, In terms of explainability/Interpretability of the variables, Decision Tree and logistic Regression do a better job.

Business Insights:

- Customers credit usage patter including payments made towards the credit in Q2 & Q3 which has been combined to create a score for each of the customers is an important variable that helps predict default.
- The History of repayment status as it stands at the beginning of the Q2 and the end of Q3 helps determine if a customer would default on their payments.
- Customer being a male has a positive impact towards default.
- Customer Education classified as “others” has a negative impact towards default.
- Age of a customer as well has a positive impact towards default.
- Customer marital status as “single” has a negative impact towards default.

Business Recommendation:

- ✓ Customer should not be allowed to spend over and above their available limit balance.
- ✓ Customer who payed duly for the month of July and if their payment amount for that month was greater than 1060 NT$ and despite that if that billed amount for September is greater than 649 NT$ and if these customers have been late in their payments in the past at least by 2 months should be considered as high Risk and can be targeted to reduce the credit available and increase in penalties for payment delays(if late by a month or more).

- ✓ Male customers who have been late by at least 2 months in their repayments as it stands in April and August , but have not been late by more than a month in the repayment in September can be categorized as medium risk and can be targeted to increase penalties for payment delays(if late by a month or more).
- ✓ Customers who are single or education as “others” can be provided with rewards point scheme or an increase in the credit available.
- ✓ Customers aged above 51 years can be categorized as high risk.