Credit Card Defaulter Detection

Arghyamalya Biswas, Roll - 201275
Suryasis Jana, Roll - 201447

April 21, 2022

(Under guidance of Dr. Amit Mitra)

Contents

1 Introduction

Banking sector is a very challenging field where various statistical data mining techniques can be used very efficiently. In our project we are analysing a very common problem of banking sector. We have collected data of a reputed bank of Taiwan. Our data contains information on age, gender, credit limit balance, payment history, bill amount history on 30000 individuals and also have the data whether they are credit defaulter or not.

Therefore, in our project our goal is to analyse the data and build a predictive model which will predict whether an individual can be a defaulter or not based on several information factors.

2 Data Description

Our study took payment data in October, 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank. Among the total 30,000 observations, 6636 observations (22.12%) are the cardholders with default payment. This research employed a binary variable – default payment (Yes = 1, No = 0), as the response variable. There are 23 variables as explanatory variables in our data set
There are 25 columns in our data set.

- ID: ID of each client
- LIMIT BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (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: Repayment status in September, 2005 (-2=No consumption,-1=pay duly in full amount, 0= The use of revolving credit (when the customer pays only the minimum balance) 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)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
2	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment
3	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
4	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
5	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
6	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
7	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
8	6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
9	7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
10	8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
11	9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
12	10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
13	11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
14	12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
15	13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
16	14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
17	15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0
18	16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0
19	17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1
20	18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0
21	19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
22	20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
23	21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0
24	22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1
25	23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1
26	24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1
27	25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0
28	26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0
29	27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1
30	28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0
31	29	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0
32	30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	0	0
33	31	230000	2	1	2	27	-1	-1	-1	-1	-1	-1	16646	17265	13266	15339	14307	36923	17270	13281	15339	14307	37292	0	0

Data

3 Data Preprocessing

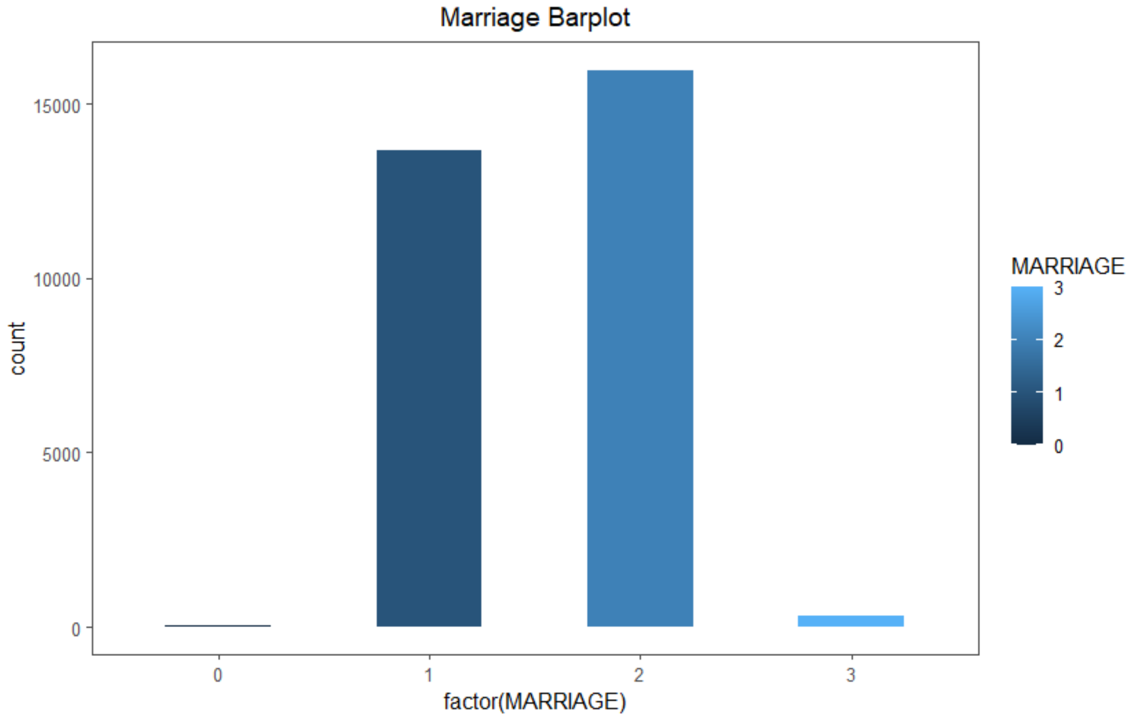
3.1 Missing Value checking

To check whether there is any missing value or not, we have counted the number of "NA" values for each of the columns and we got that for each column the number of missing value is 0. So there is no missing value.

```
> sapply(credit, function(x) sum(is.na(x)))
                        ID                  LIMIT_BAL                        SEX 
                         0                          0                          0 
                 EDUCATION                   MARRIAGE                        AGE 
                         0                          0                          0 
                     PAY_0                      PAY_2                      PAY_3 
                         0                          0                          0 
                     PAY_4                      PAY_5                      PAY_6 
                         0                          0                          0 
                 BILL_AMT1                  BILL_AMT2                  BILL_AMT3 
                         0                          0                          0 
                 BILL_AMT4                  BILL_AMT5                  BILL_AMT6 
                         0                          0                          0 
                  PAY_AMT1                   PAY_AMT2                   PAY_AMT3 
                         0                          0                          0 
                  PAY_AMT4                   PAY_AMT5                   PAY_AMT6 
                         0                          0                          0 
default.payment.next.month 
                         0 
```

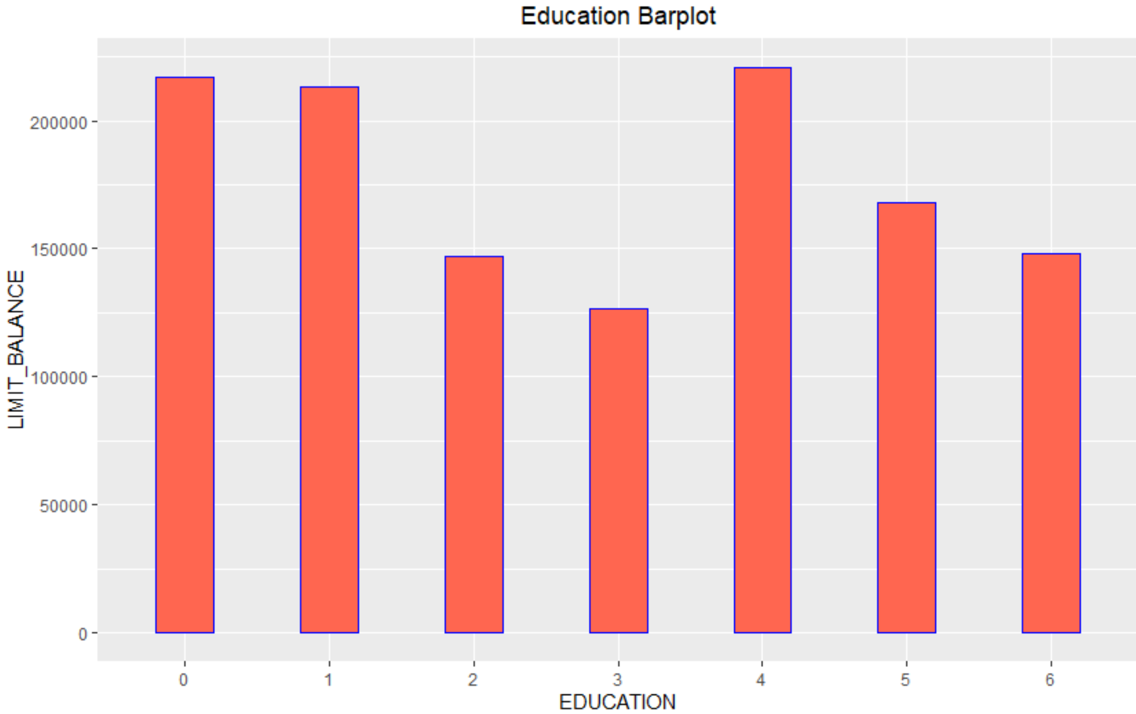
3.2 Category Reassignment for some variables

We have noted that for some of the categorical covariates there are some misleading entries. For example, in the data description it was mentioned that there are 3 categories of MARRIAGE viz. 1,2,3. But in the data there were 54 cases where the value corresponding to MARRIAGE is 0.

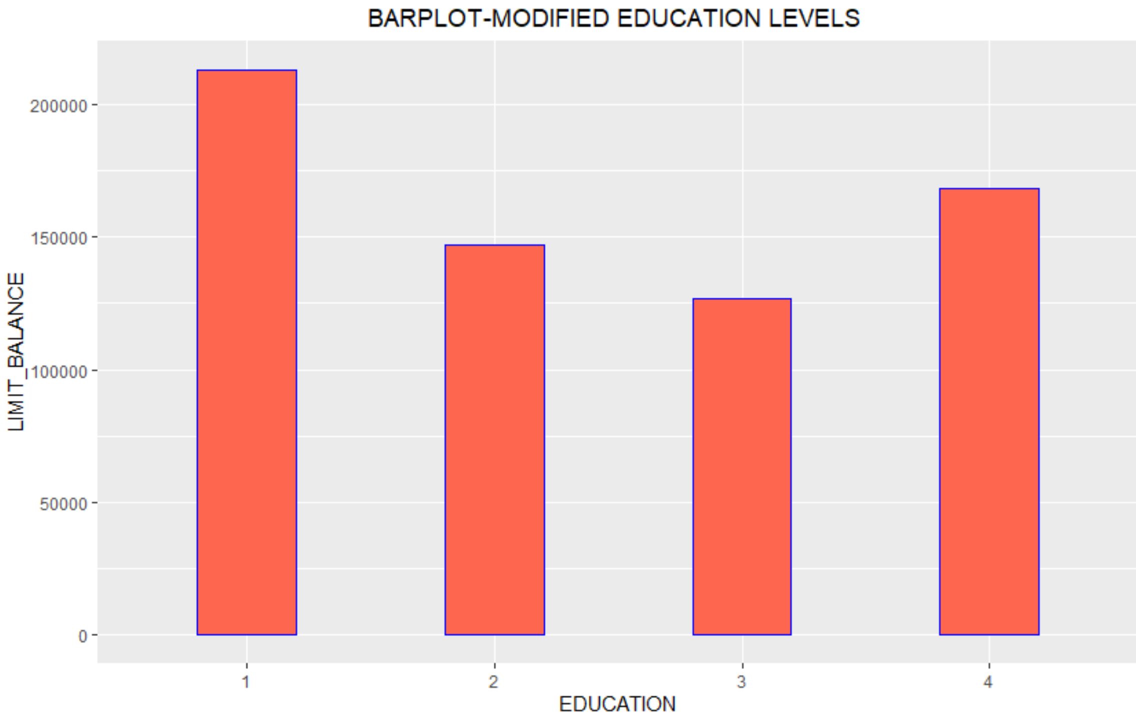


It would be very reasonable to assume that in a typical society of credit card users, the average age for getting married can be around 27. So here we replace the value 0 with 1 for MARRIAGE for the individuals (clients) having age more than 27 and replace with 2 for the clients having age is less than or equal to 27.

Education is an important factor for credit card limit balance Now we perform the reassignment of the categories for the education level factor. We observe the categories 4, 5, 6 are technically same [1=graduate school, 2=university, 3=high school, 4=others, 5=un known, 6=unknown]. Moreover when we plot the average limit balance with respect to the different categories of Education level, we note the existence of an additional category 0, which is not mentioned in the data description. So we intend to club it with the category having the most likely average limit balance.

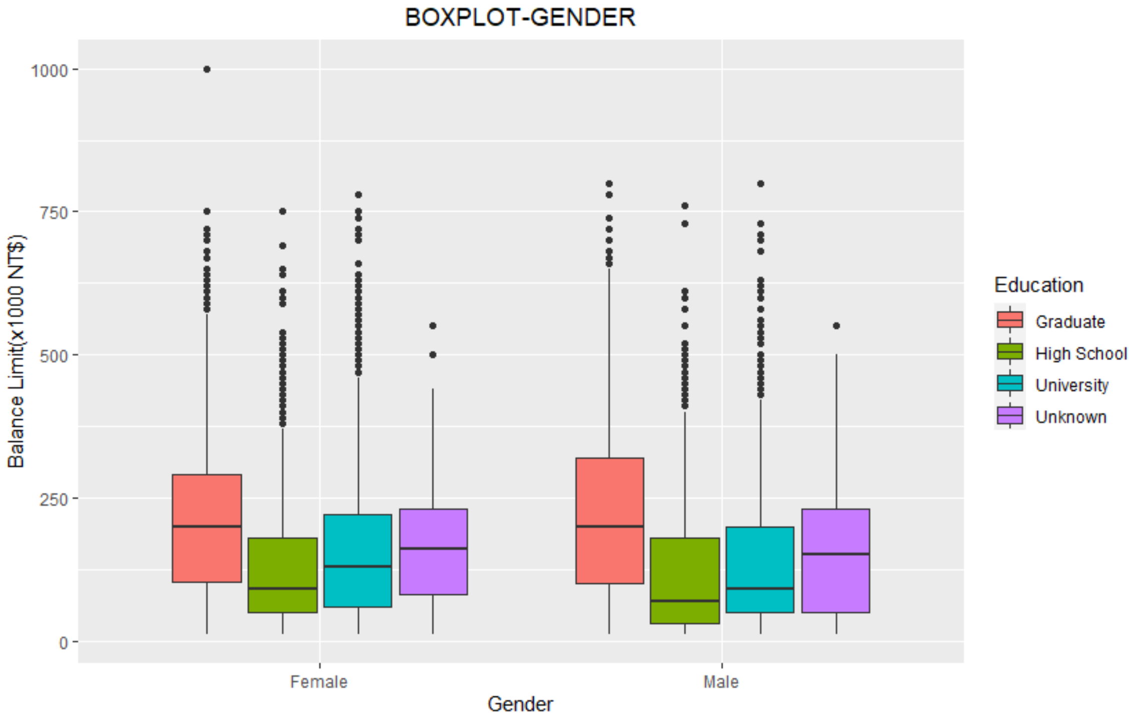


It may be justified to assume that the credit limit balance is proportional to the educational level, hence we replace the category 0 and 4 with 1 as 0 and 4 have the more or less same average limit balance. Similarly we have replaced category 6 with category 2 and renamed the category 5 as category 4.

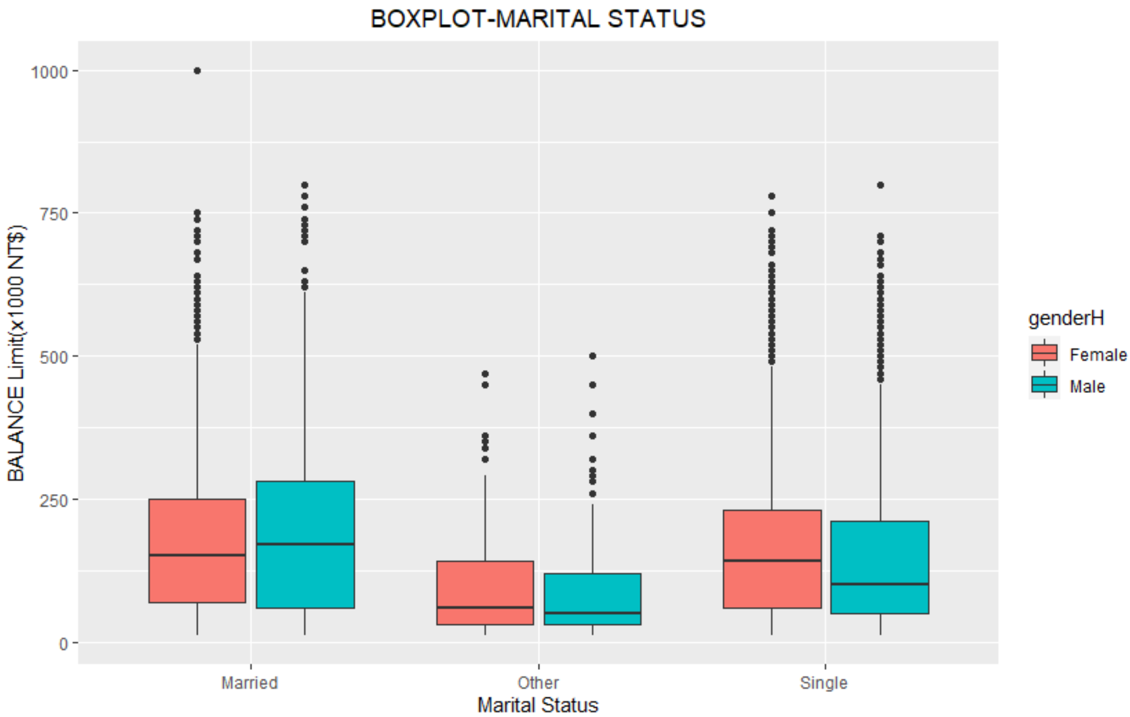


3.3 Outlier Detection

Since outliers may highly affect a predictive analysis so we intend to detect the presence of outliers in our data through boxplot visualizations. In the following diagram we have generated the boxplot for each gender and education categories.



Also the same plot for each gender and marital status categories are given below.

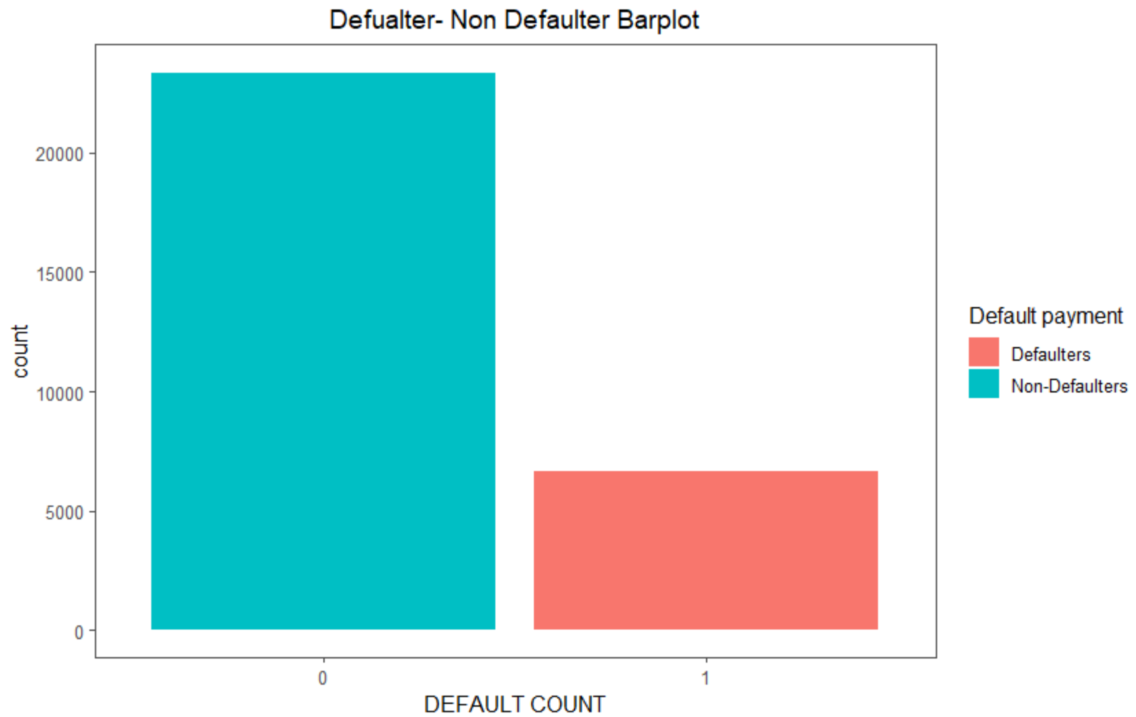


The data points with credit limit 10^6 NT-Dollar has been detected as potential outlier in our analysis. So, we have removed them and performed the subsequent analysis with the modified data.

4 Visualization

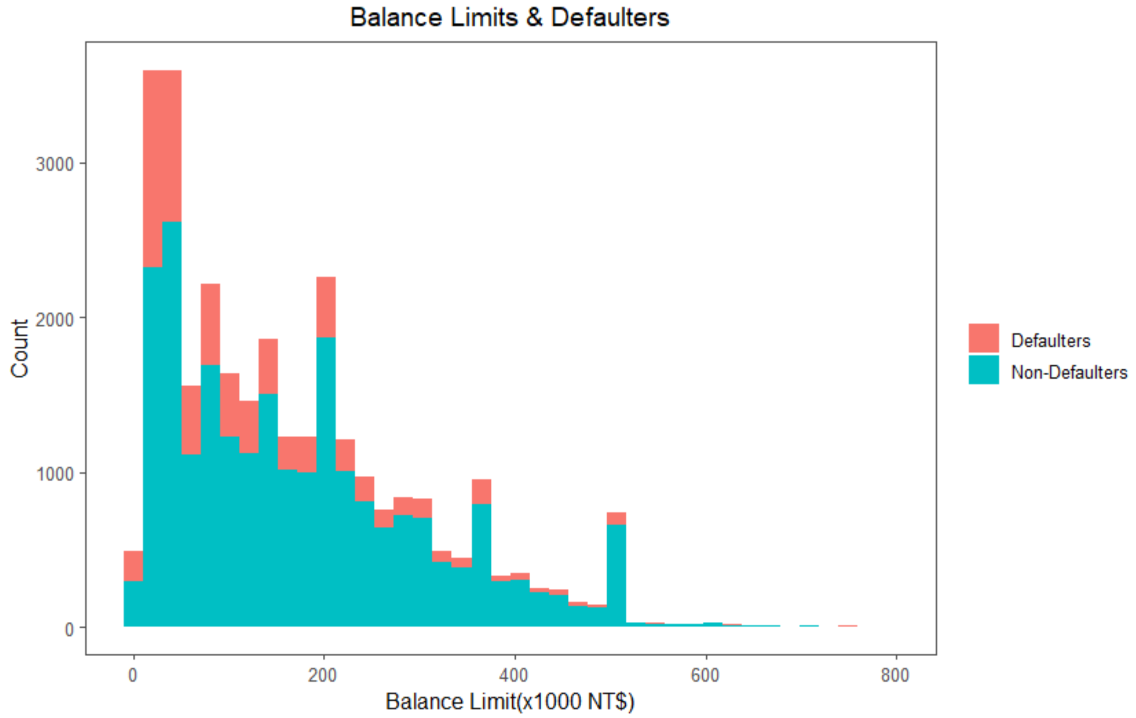
4.1 Response Visualization

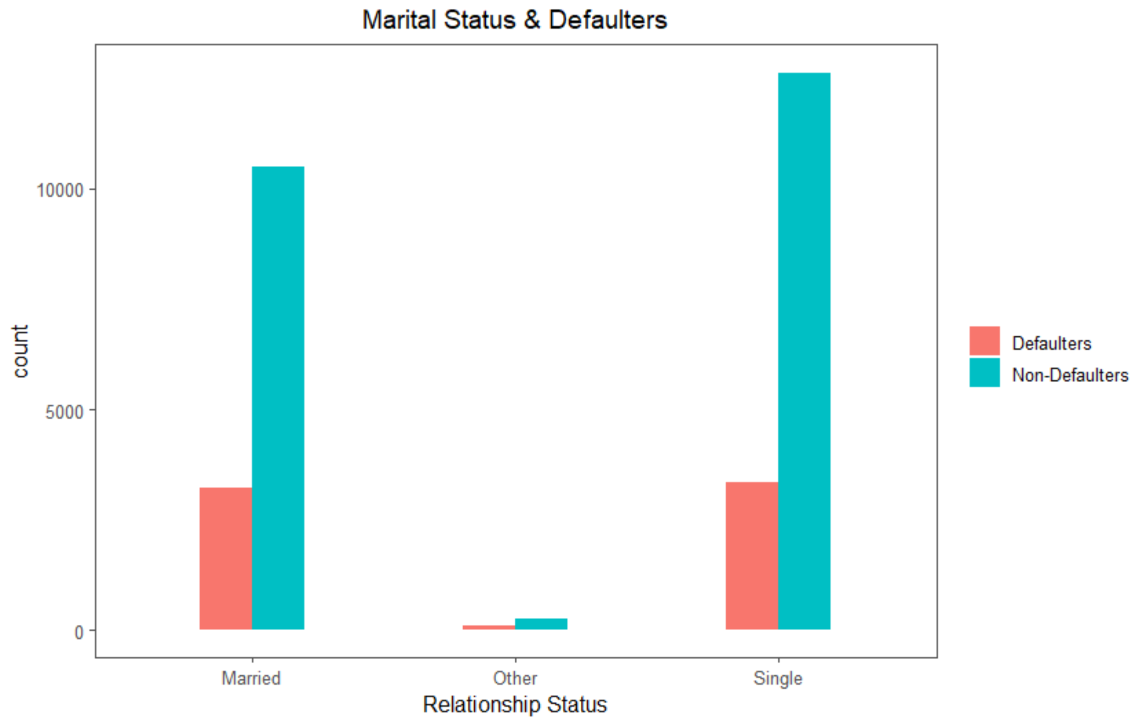
Here we are interested to observe the ratio of non-defaulter and defaulter in our data set. So we have obtained the barplot of the response variable(default.payment.next.month), which is the indicator variable showing a particular ID is defaulter or not.



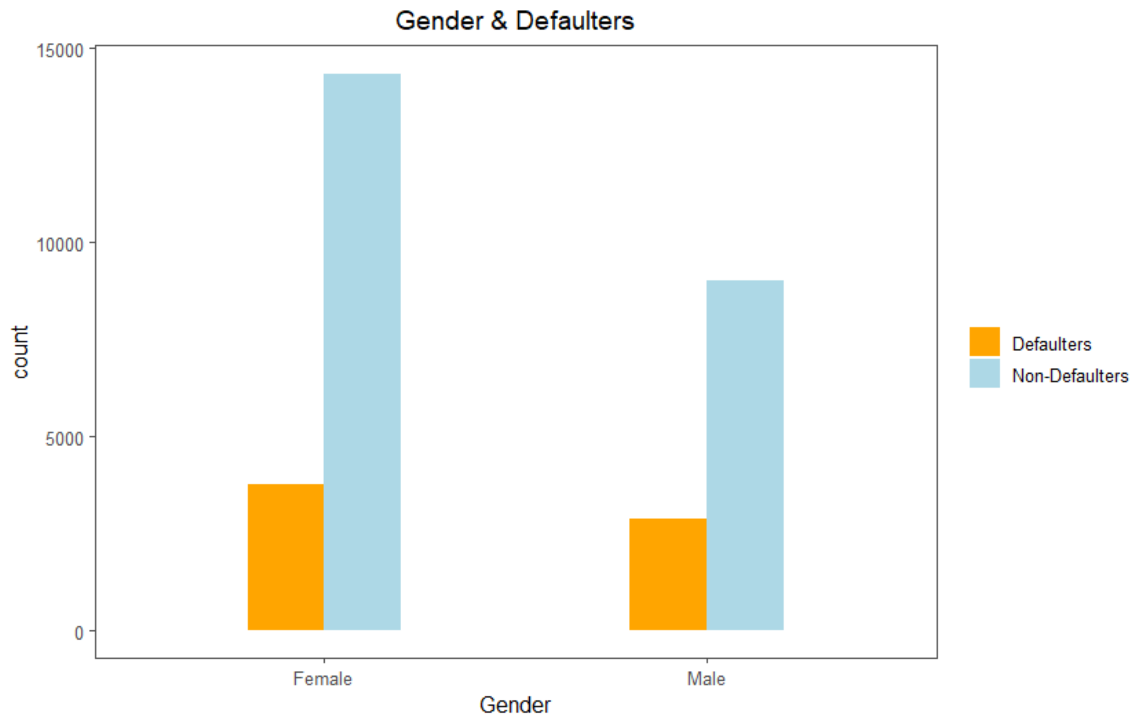
So, from the above plot we observe more than 80% of the individuals in our data set are non-defaulters.

The following graph shows the distribution of credit limit balance of defaulter and non-defaulter individuals.



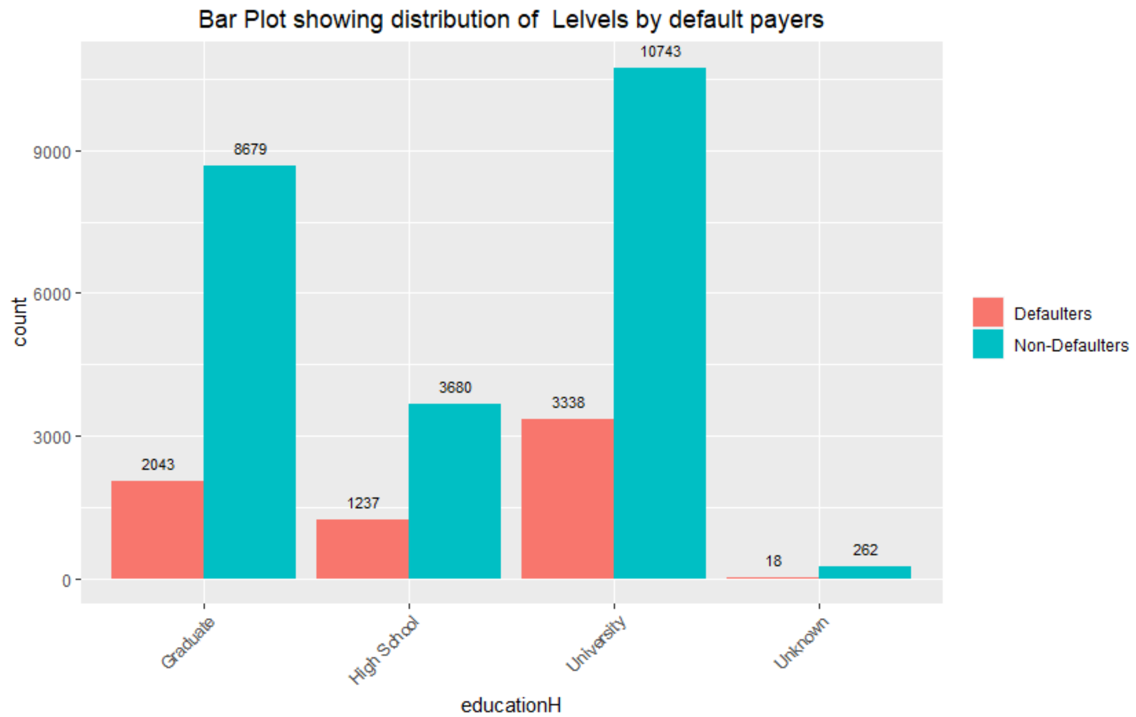


The above plot simply shows the proportion of defaulters and non defaulters for different marital status. Note that, the proportion of non-defaulters is more for the singles than the marrieds.



This plot gives us an indication that how proportion of defaulters and non-defaulters varies among genders. We see that among females the proportion of non defaulters is more

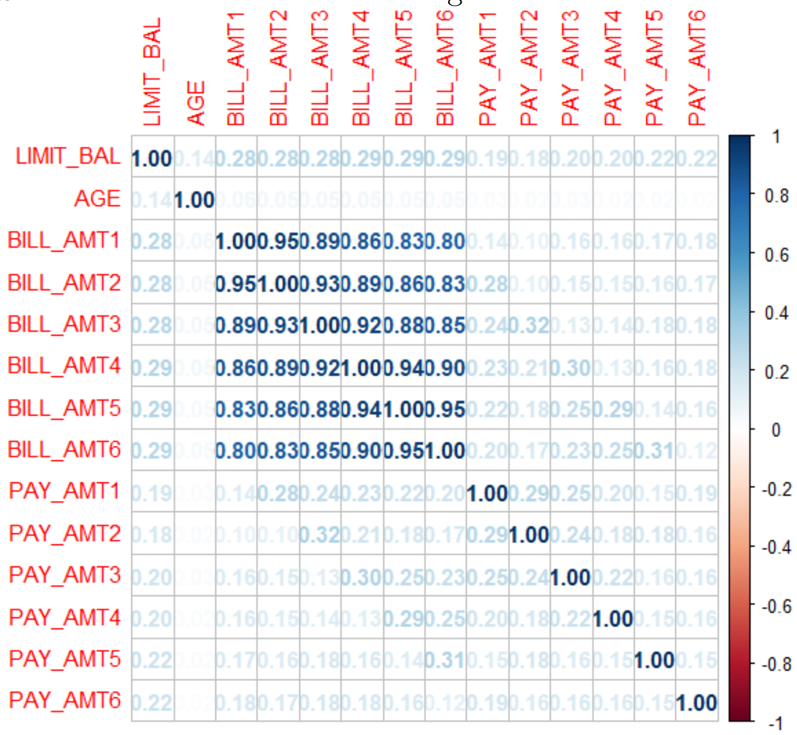
The above plot shows the proportions of defaulters and non-defaulters for various payment status(as mentioned in the data description). Here we observe that for payment status -2,-1,0 the proportion of non-defaulters are significantly higher than defaulters. And for payment status 2, 3 the proportion of defaulters are significantly higher than non-defaulters.



The above plot shows the proportions of defaulters and non-defaulters for the education levels.

4.2 Covariate Visualization

We consider the variables Limit balence, Age, Bill amounts, Pay amounts as our predictors. Now we proceed to see whether the predictors are correlated among themselves. So we have obtained the following correlation matrix.



From the plot we observe that covariates BILL-AMT1,...,BILL-AMT6 are highly correlated with each other. And among the other covariates there is no such high correlation. So, because of this instead of using BILL-AMNT1 to BILL-AMNT6 it is sufficient to use BILL-AMNT1 in our analysis.

5 Predictive Models

We have splitted our whole data set into 2 parts: Training data set(containing 70% of the observations) and Test data set(containing 30% of the observations). Then we have trained our predictive models on the training data set and validated those models on the test data set.

5.1 Logistic Regression

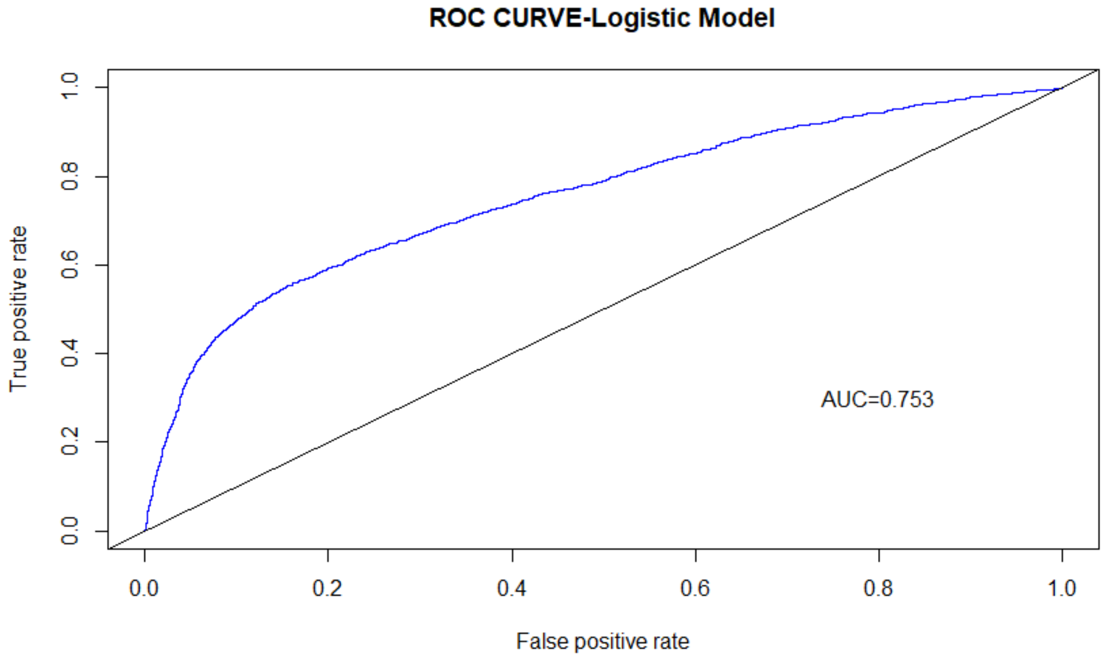
Since our response variable is binary, so as a predictive model we have fitted the Logistic Regression in the data. This model fits the log-odds of the probability of the event as a linear function of the covariates.
Note that the covariates "SEX", "EDUCATION", "MARRIAGE","PAY-0", "PAY-2", "PAY-3","PAY-4","PAY-5", "PAY-6", are categorical variables. We have estimated the parameters of the logistic regression model using maximum likelihood method.

We have got the accuracy measure as 80.36%. The following result shows the accuracy results of our fitting.

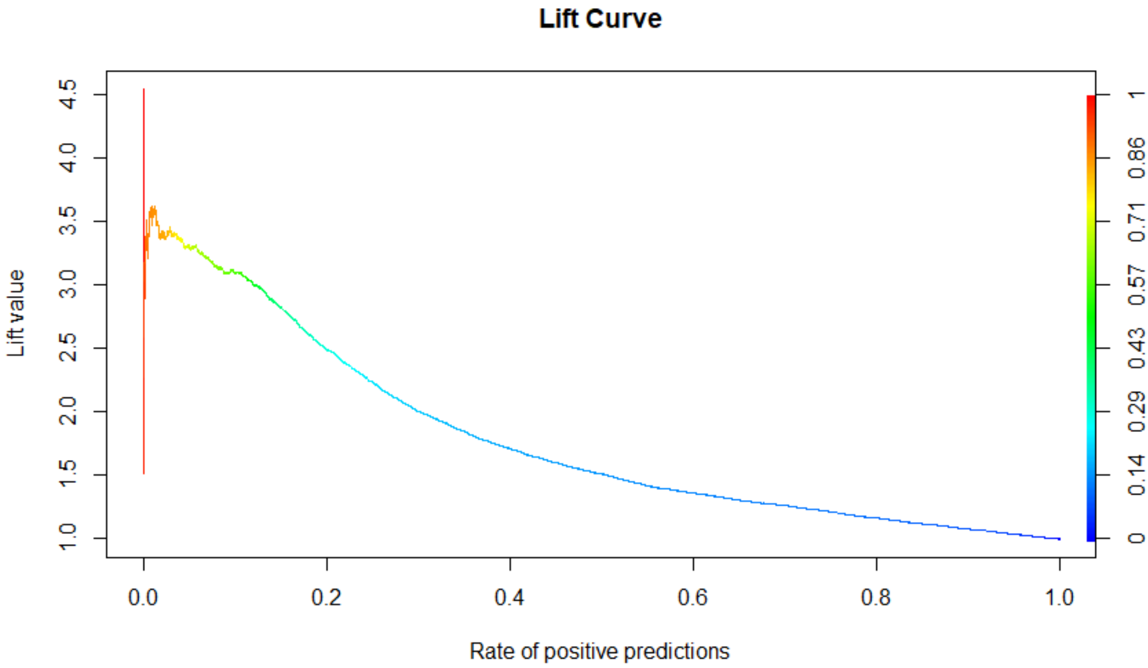
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6275  743
         1 1025  957

               Accuracy : 0.8036
                 95% CI : (0.7952, 0.8117)
    No Information Rate : 0.8111
    P-Value [Acc > NIR] : 0.967

                  Kappa : 0.3973

 Mcnemar's Test P-Value : 2.343e-11

            Sensitivity : 0.5629
            Specificity : 0.8596
         Pos Pred Value : 0.4828
         Neg Pred Value : 0.8941
              Precision : 0.4828
                 Recall : 0.5629
                     F1 : 0.5198
             Prevalence : 0.1889
         Detection Rate : 0.1063
   Detection Prevalence : 0.2202
      Balanced Accuracy : 0.7113

       'Positive' Class : 1
```

Here, for the prediction purpose, we have taken the threshold value of the estimated probability of an observation belonging to class "1" ($\hat{P}(\pi_1|x)$) as 0.3. Which means if, for an individual client the probability that, he is going to be defaulter in the next month , then we are classifying him as a defaulter. This is because we don't want to take risk in identifying a defaulter.

Model evaluation curves:



We observe that sensitivity (or True Positivity Rate) is 0.56 where as Specificity (True Negativity Rate) is 0.86, so the logistic regression model based classifier incorrectly classifies the defaulter clients more often which has affected the above Receiver Operating Characteristics(ROC) curve. The more the curve is shifted towards top-left of the ROC space the better is the classifier. We, now observe the AUC (Area under Curve) below for the logistic model.

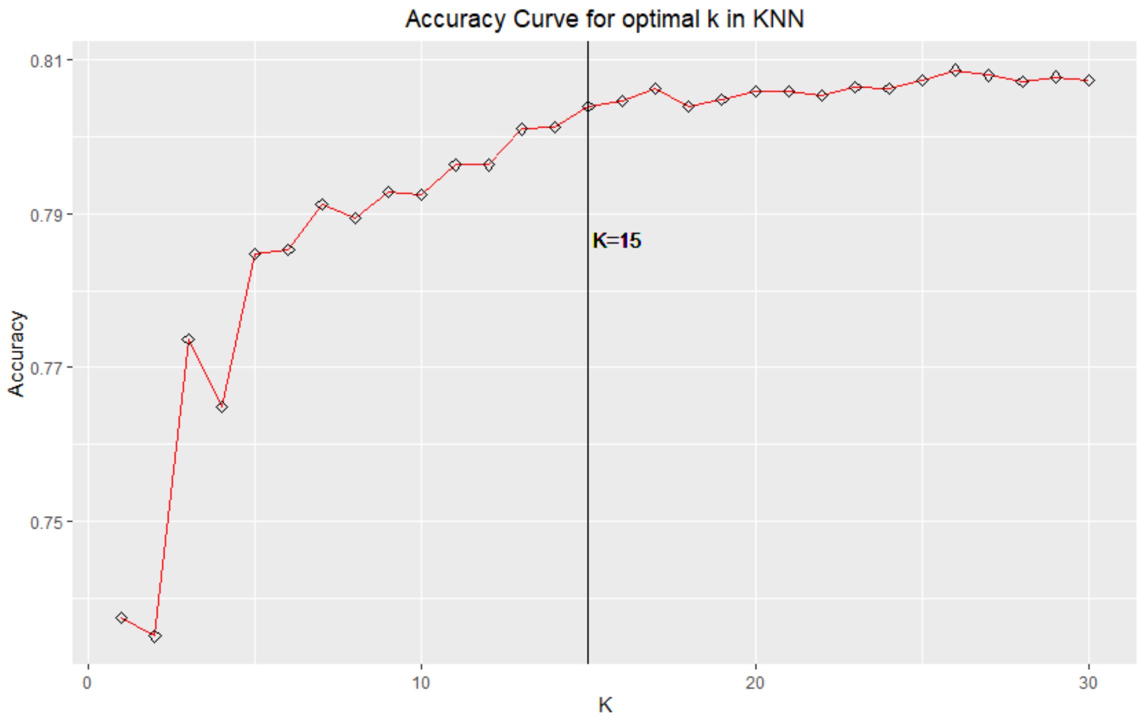


The lift curve suggests that the fitted predictive model is fine. The lift curve plots lifts on the Y-axis and Rate of positive predictions or Sample proportions on the X-axis.

5.2 K-NN Classifier

Now we also have fitted K-NN Classifier in our data as a predictive model. The main concept of K-NN classifier is the following:
Let $\mathcal{L} = \{(x_i, y_i) : i = 1, 2, \ldots, n\}$ be our learning sample, where x_i is the covariate vector for i^{th} observation and y_i is the binary response variable taking only value 0 or 1. Then for a new observation with input vector x, in order to classify it to 0 or 1 we consider the k neighbourhood of x and then apply majority voting rule in that neighbourhood and to the winner group we classify the observation.

To, select the optimal value of K, we have obtained the the Accuracy-plot where we have plotted the accuracy measures for various values of k.



Here the accuracy measure is being used as

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ no\ of\ observations}$$

From the above Accuracy-plot we see that after K = 15 the accuracy is not increasing significantly. So, in our analysis we have taken K = 15.
The confusion matrix of the our fitted K-NN model is the following:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7172  128
         1 1598  102

               Accuracy : 0.8082
                 95% CI : (0.7999, 0.8163)
    No Information Rate : 0.9744
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0635

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

            Sensitivity : 0.44348
            Specificity : 0.81779
         Pos Pred Value : 0.06000
         Neg Pred Value : 0.98247
              Precision : 0.06000
                 Recall : 0.44348
                     F1 : 0.10570
             Prevalence : 0.02556
         Detection Rate : 0.01133
   Detection Prevalence : 0.18889
      Balanced Accuracy : 0.63063

       'Positive' Class : 1
```

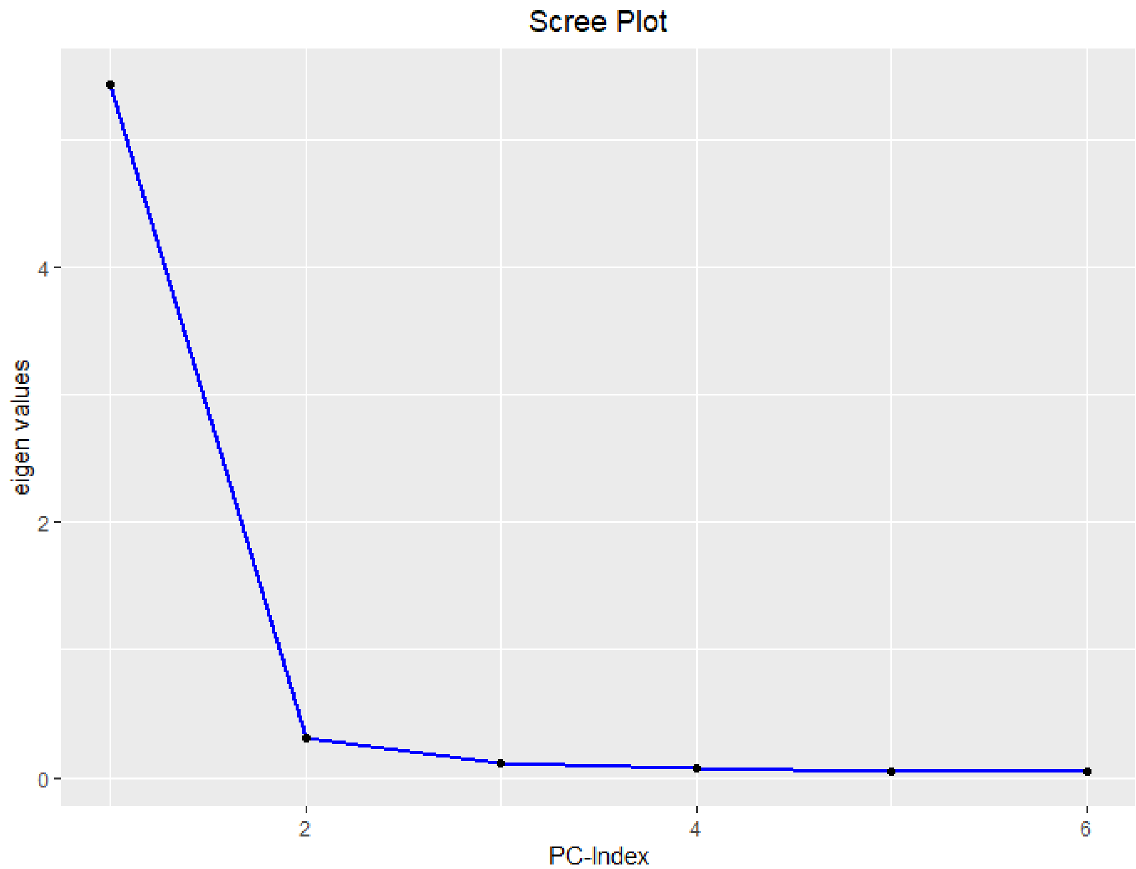
From the above result we observe that Accuracy measure is 80.82%. Also, note that sensitivity is 0.44348, Specificity is 0.81779 and F1 score is 0.10570.

5.3 Principal Component Analysis(PCA)

In the correlation matrix we observed that there is high correlation among the covariates BILL_AMT1,...,BILL_AMT6. So, now in order to reduce the dimension we proceed to apply PCA. Here we have applied PCA on the statndardized variables. The result of the analysis is shown below:

```
> A=credit[,13:18]   # Contains BillAMT-1 to BillAMT-6
> PCA = prcomp(A,scale=T,center = T)
> PCA
Standard deviations (1, .., p=6):
[1] 2.3309473 0.5531005 0.3341400 0.2591488 0.2039773 0.2008737

Rotation (n x k) = (6 x 6):
                PC1        PC2         PC3         PC4        PC5        PC6
BILL_AMT1 0.4008762  0.5364617 -0.45738377  0.22423436 -0.4873929 -0.2333973
BILL_AMT2 0.4105554  0.4288167 -0.11594854 -0.09036302  0.7043317  0.3603719
BILL_AMT3 0.4120351  0.1755554  0.62220844 -0.55452135 -0.1771723 -0.2708876
BILL_AMT4 0.4147615 -0.1807669  0.43976201  0.59180341 -0.2152428  0.4531500
BILL_AMT5 0.4102913 -0.4301220 -0.08032282  0.27540617  0.3772215 -0.6496613
BILL_AMT6 0.4007539 -0.5289838 -0.43631446 -0.45604792 -0.2152893  0.3389292
> R = cor(A)   # We are applying PCA on the standardized variables
> eval = eigen(R)$values
> cumsum(eval)/sum(eval)
[1] 0.9055525 0.9565392 0.9751475 0.9863405 0.9932750 1.0000000
```

So, from the above result we see that the first principal component explains almost 90% variation of the data consisting of BILL_AMT1.....BILL_AMT6; So, in our subsequent analysis instead of using these 6 variables, we shall be using the first principal component.

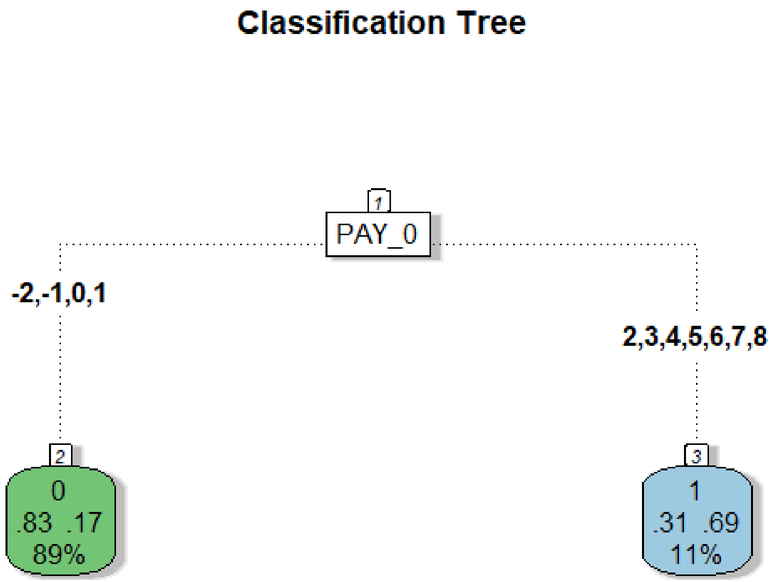


Let B_i^* denotes the standardization of $BILL\_AMTi$, $i = 1, 2, \ldots, 6$. Then the first principal component is given by:

$$PC1 = 0.401B_1^* + 0.411B_2^* + 0.412B_3^* + 0.414B_4^* + 0.410B_5^* + 0.401B_6^*$$

5.4 Classification Tree

We have applied Classification Tree algorithm to our data set. Since our response variable is binary, so we have a two class problem. The obtained classification tree is the following:

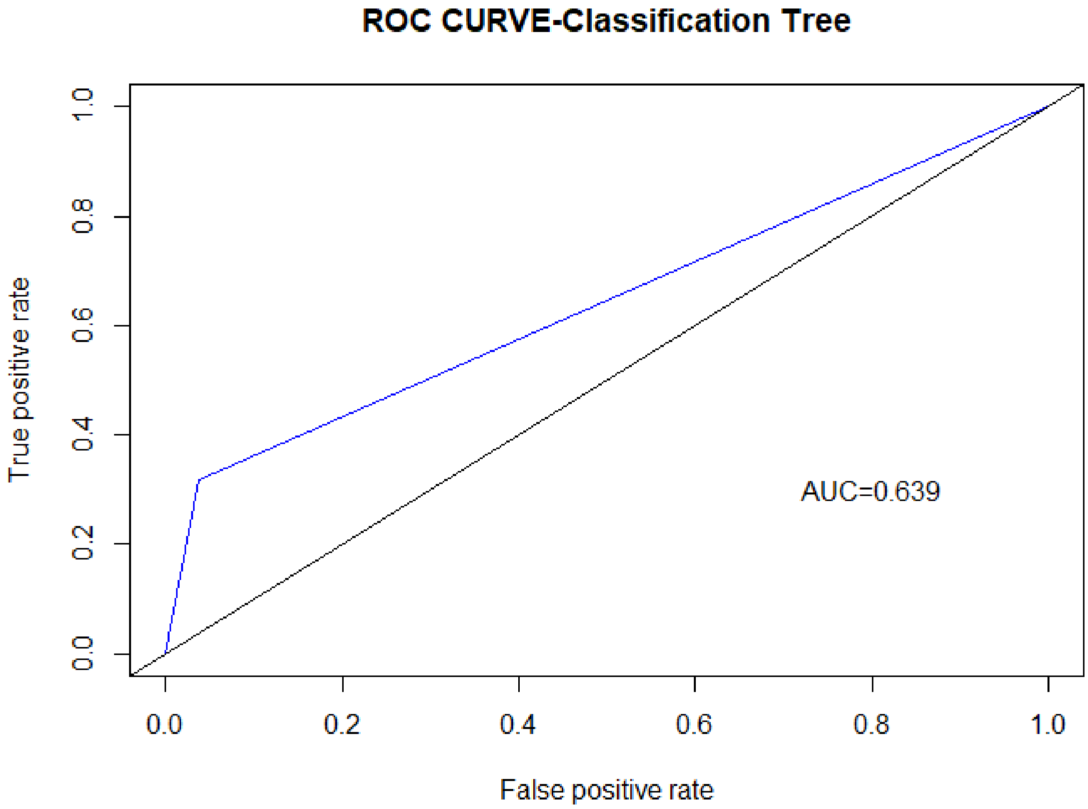


Here we see that PAY_0 is the split variable at the root node. And if it's value is ≥ 2 then, we classify the observation as a defaulter, otherwise we classify it as a non-defaulter.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6721  275
         1 1340  664

               Accuracy : 0.8206
                 95% CI : (0.8125, 0.8284)
    No Information Rate : 0.8957
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3604

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

            Sensitivity : 0.70714
            Specificity : 0.83377
         Pos Pred Value : 0.33134
         Neg Pred Value : 0.96069
              Precision : 0.33134
                 Recall : 0.70714
                     F1 : 0.45124
             Prevalence : 0.10433
         Detection Rate : 0.07378
   Detection Prevalence : 0.22267
      Balanced Accuracy : 0.77045

       'Positive' Class : 1
```

We observe that the accuracy measure is coming as almost 82%. So, our fitting is good. Also both the sensitivity and specificity are as high as 0.707 and 0.834. Also the F1 score is 0.451.

Next we have obtained the following ROC plot



As, shown by the plot, the area under the curve is 0.693, which is not a bad result.

5.5 Random Forest Classifier

Next, we have applied Random Forest classifier to our data, which is an ensemble learning model. The main concept behind Random Forest is that:

First, we have to take several bootstrap samples from our given data set.
Then, we have to apply a classification tree algorithm to each such bootstrap sampled data sets. So, now we have several number of classifiers.
Now, to classify a new observation we input the data to each classifier and collect the result from each of them.
Then on the collected results we apply Majority Voting Rule.

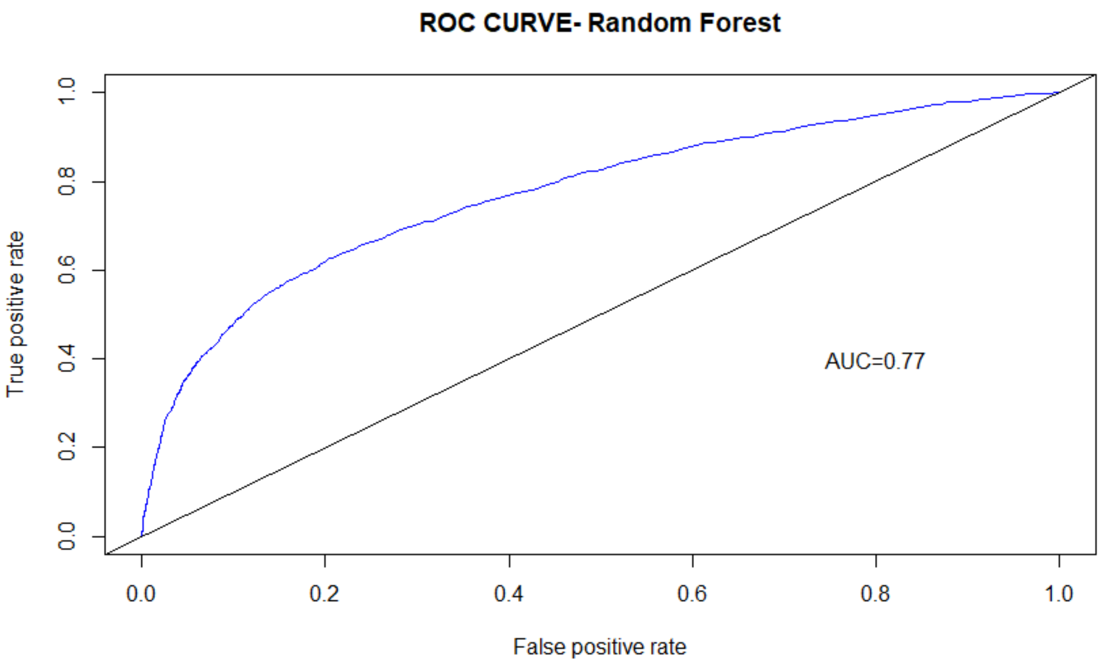
After applying Random Forest Classifier to our data we have got the following accuracy results:

```
Call:
 randomForest(formula = default.payment.next.month ~ ., data = train.tree,      method = "class")
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 18.02%
Confusion matrix:
      0    1 class.error
0 15452  876  0.05365017
1  2909 1763  0.62264555
```

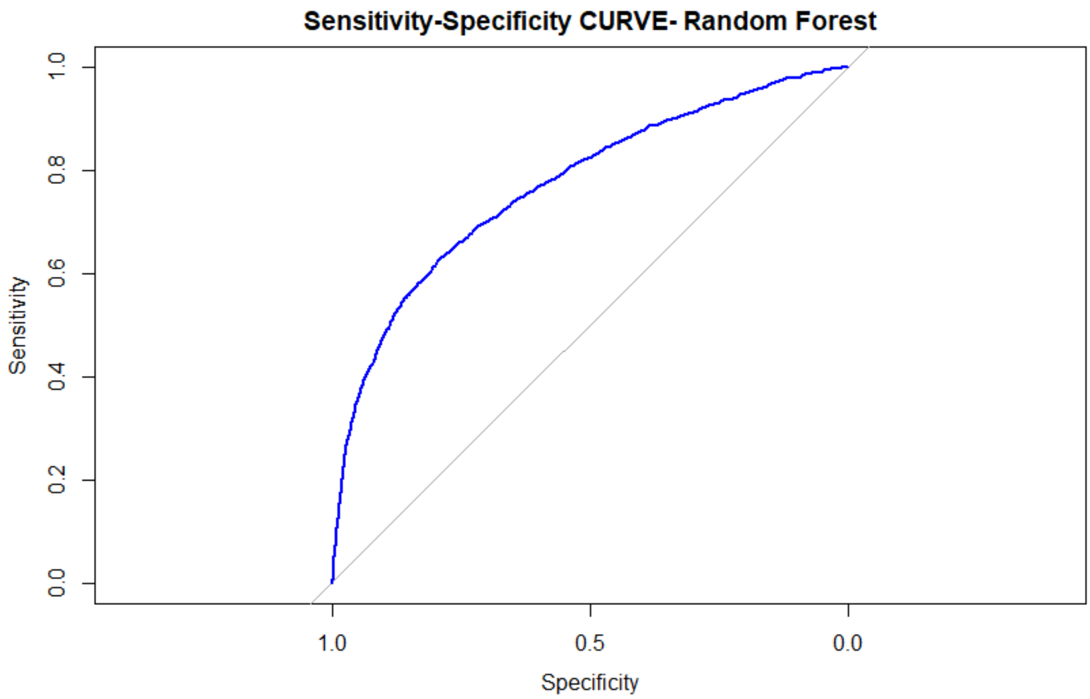
Now using the test data set we have obtained the following accuracy results.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6682  354
         1 1251  713

               Accuracy : 0.8217
                 95% CI : (0.8136, 0.8295)
    No Information Rate : 0.8814
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3743

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

            Sensitivity : 0.66823
            Specificity : 0.84230
         Pos Pred Value : 0.36303
         Neg Pred Value : 0.94969
              Precision : 0.36303
                 Recall : 0.66823
                     F1 : 0.47047
             Prevalence : 0.11856
         Detection Rate : 0.07922
   Detection Prevalence : 0.21822
      Balanced Accuracy : 0.75527

       'Positive' Class : 1
```

From this result we observe that the accuracy measure is 82.17%. But the value of sensitivity we are getting as 0.668. Also here the specificity is 0.842 and the F1 score is 0.47047.



From the above ROC plot we can see that the area under the curve is 0.77, which is more or less good.



The above diagram we plot the Specificity in the x axis, where Specificity is 1 - FPR (False Positivity Rate).

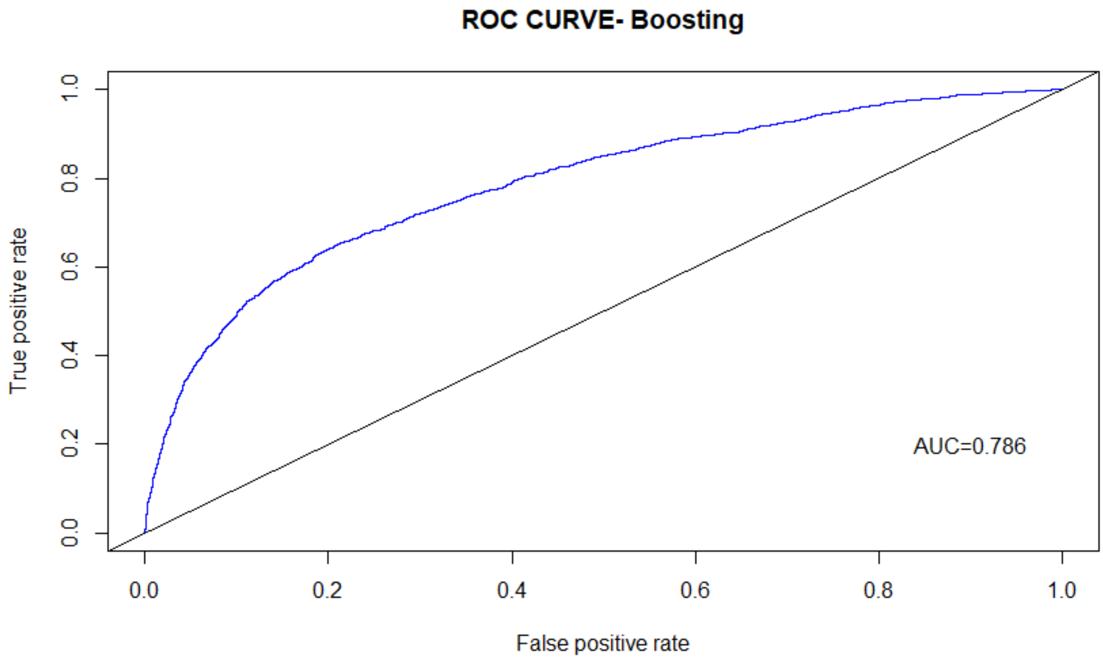
5.6 Boosting

Next, we have used another ensemble learning model "Boosting" in our data. The following figure is the accuracy result:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6707 1276
         1  329  688

               Accuracy : 0.8217
                 95% CI : (0.8136, 0.8295)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3674

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.35031
            Specificity : 0.95324
         Pos Pred Value : 0.67650
         Neg Pred Value : 0.84016
              Precision : 0.67650
                 Recall : 0.35031
                     F1 : 0.46159
             Prevalence : 0.21822
         Detection Rate : 0.07644
   Detection Prevalence : 0.11300
      Balanced Accuracy : 0.65177

       'Positive' Class : 1
```

From the above result we observe that the sensitivity is 0.35, specificity is 0.95 and F1 score is 0.46. And the accuracy measure is coming out to be 82.17%.

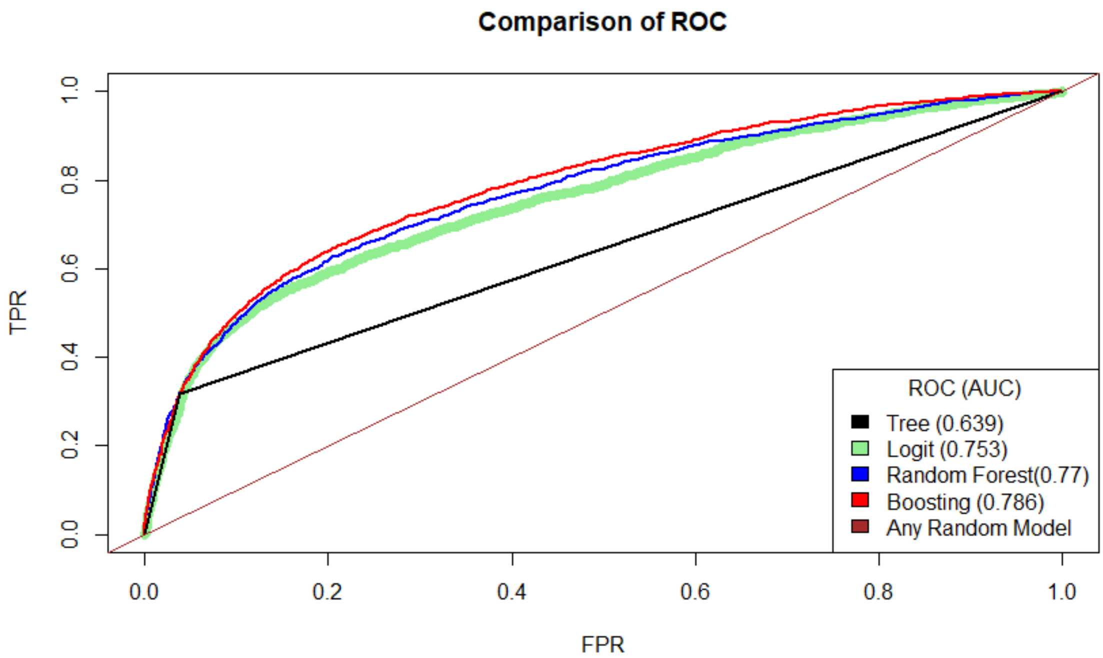
The following plot is the ROC curve for Boosting:



In the above plot the AUC is is 0.786, which is a good indication.

6 Conclusion

In the following diagram we have plotted the ROC curves for various models discussed above.



We observe that for Boosting the Area Under the ROC curve is the highest, with the AUC = 0.786.

It is interesting to observe that the F1 score for logistic model is 0.5198, which is better than other models for the given data set. Moreover it is to be noted that the classification tree has lower AUC compared to others although having good accuracy value = 82.06 and far better sensitivity value = 0.707.

Actually, on an average all of these models have low sensitivity and high specificity, which can be attributed to the fact that original data has a very low ratio of defaulters.

If someone wants to emphasize on detecting the defaulters mostly then the Classification tree would be a valid option here.

At last we may conclude that Random Forest model would be a better choice where as the K-NN Classifier would be the worst choice for the given dataset.

7 References

1. We have collected data from Kaggle.
2. Class notes of Dr. Amit Mitra

8 Acknowledgement

Project is like a bridge between the theoretical learning and practical working. With this will, we all started this project. First of all, We are feeling oblige in taking the opportunity to express our sincere thanks and gratitude to our professor **Dr. Amit Mitra** who gave us the golden opportunity to execute this Project and also helped us to complete it. We also feel delighted to thank all our seniors and classmates who are also somehow responsible for the successful completion of this project.