Predicting Default of Credit Card Clients

Ekaagar Singh Hara- G01050783
Srijan Yenumula- G01045321
Adithya Job- G01041118
OR-568

Table of Contents

Introduction

The banks have become more rigorous in issuing credit cards and in evaluation of the applicant's credit worthiness. The lending based on credit worthiness involves large number of decision making and banks rely on models rather than human discretion. The aim of this project is to formulate a model with which we can accurately predict whether an existing credit user will be a defaulter on his new credit card. We look forward to discovering interesting facts hidden in the dataset along the path of the project i.e., the factors that might not seem very important but play a major role in the classification.

Data Description

The data set selected is from UCI Machine Learning Repository. This data is of banking customers in Taiwan and it records if a customer has defaulted in the past or not. The characteristics of the data is multivariate and records 30,000 instances. The total number of predictors are twenty-four and its characteristics are categorical or numeric.

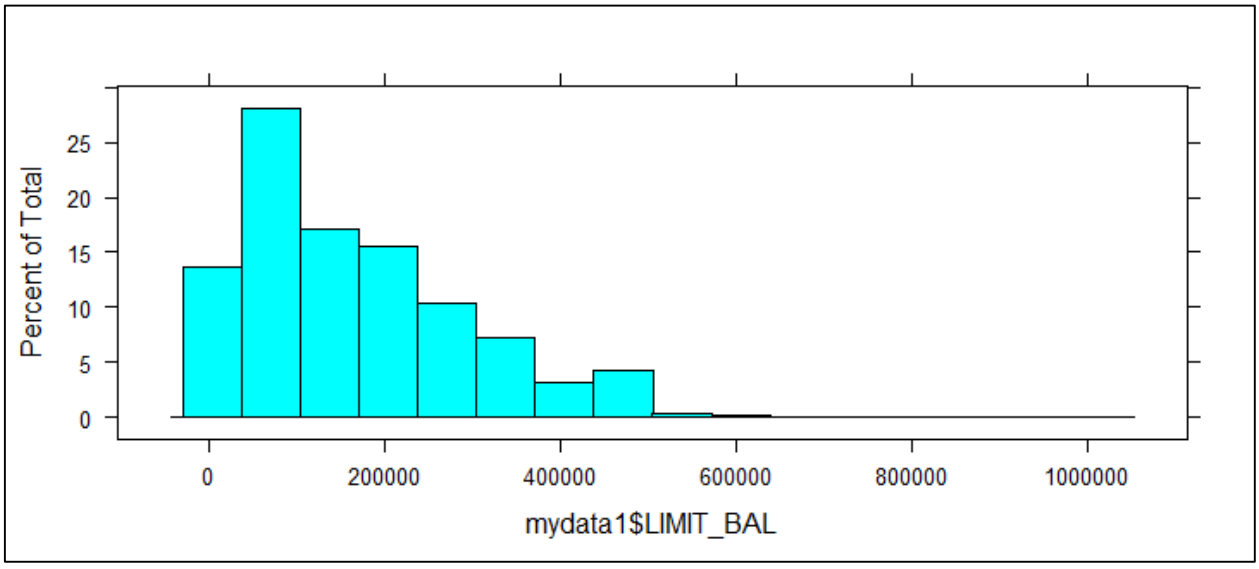
Number of predictors: 23 (3-Categorical, 20-Numeric)

Response Variable: Whether defaulted on payment or not (Categorical)

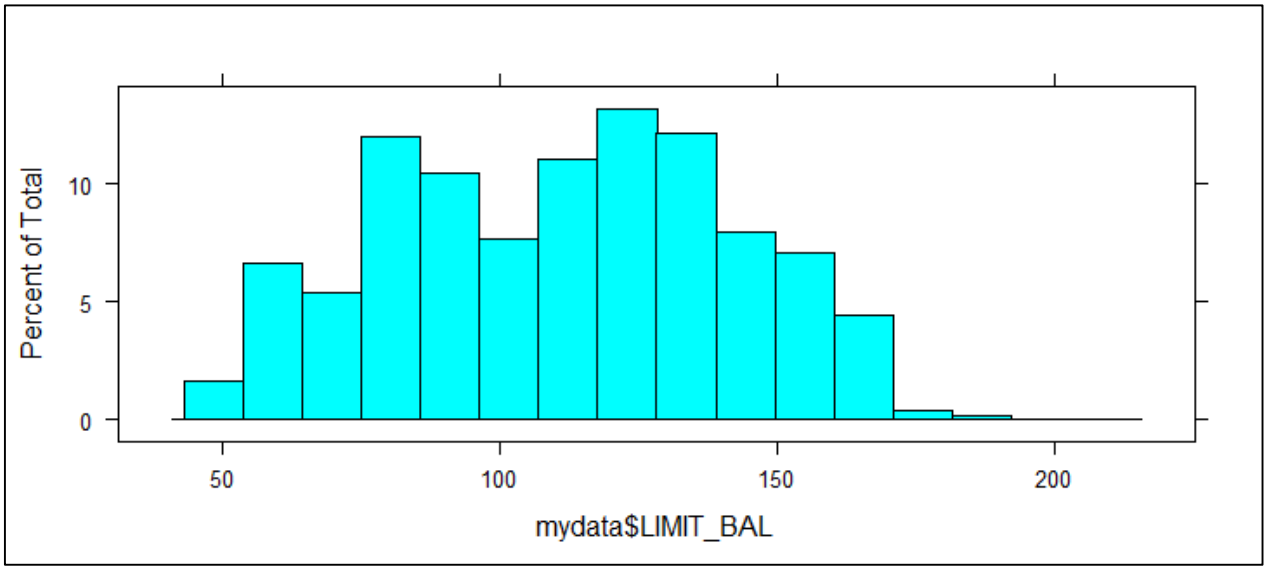
Variables	Comment	Type of Data
Response Variable	Default Payment	Categorical (Yes = 1, No = 0)
LIMIT_BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.	Numeric
SEX	Gender	Categorical (1 = male; 2 = female)
EDUCATION	Educational Qualification of the customer	Categorical (1 = graduate school; 2 = university; 3 = high school; 4 = others)
MARRIAGE:	Marital Status	Categorical (1 = married; 2 = single; 3 = others)
AGE	Age of customers in years	Numeric
PAY_0 to PAY_6	These six predictors show the history of past payment for a customer. The past monthly payment records are from September to April, 2005.	Numeric (The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months and similar scale upto 9
BILL_AMT1 to BILL_AMT6	Amount of bill statement (NT dollar) from September to April 2005.	Numeric

Data Pre-processing

We conducted PCA on our dataset but the results that we got were not appreciable, so we stopped that approach. The next approach we tried was Box-Cox transformation and with it, we were able to reduce the skewness of our predictors and thereby we made the continuous predictors normalized. For example, the LIMIT_BAL predictor before box cox transformation had a skewness of 0.9927492, which after transformation became -0.02251852. Similarly, all the other predicators were normalized.

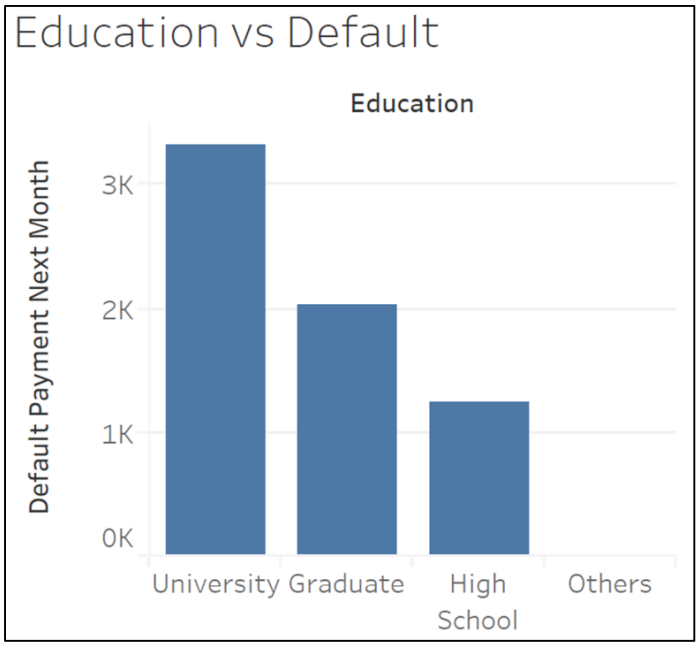


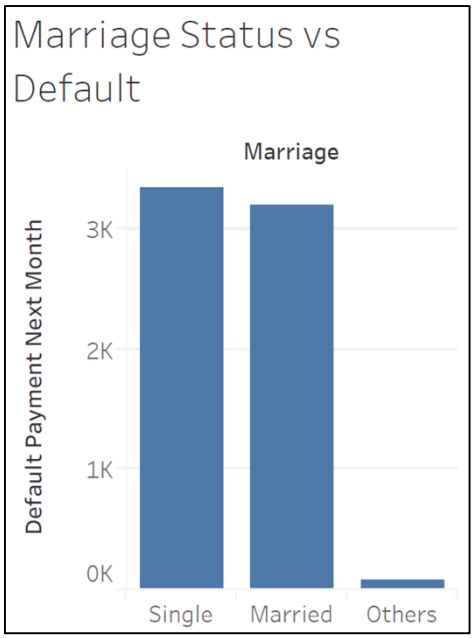
After transformation:



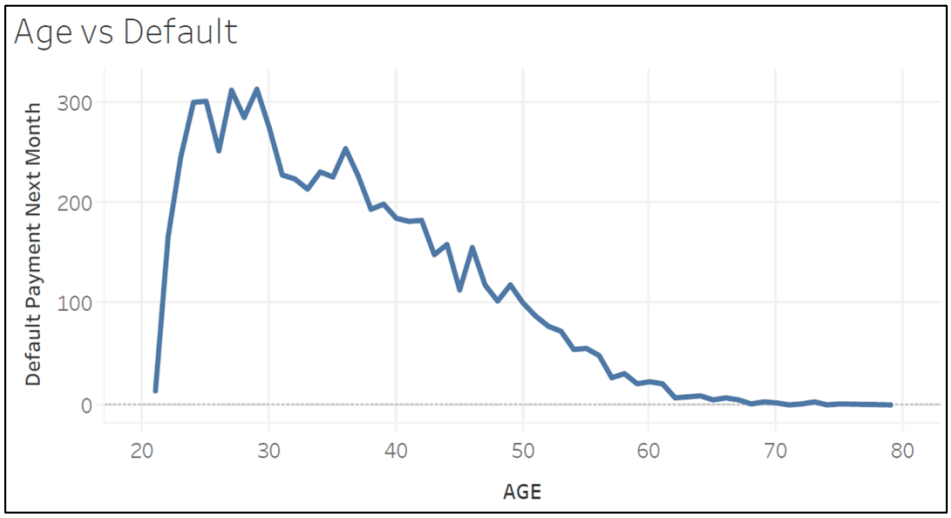
Data Exploration

Using tableau, we found some basic insights of our data. Below are the graphs that we generated using tableau:

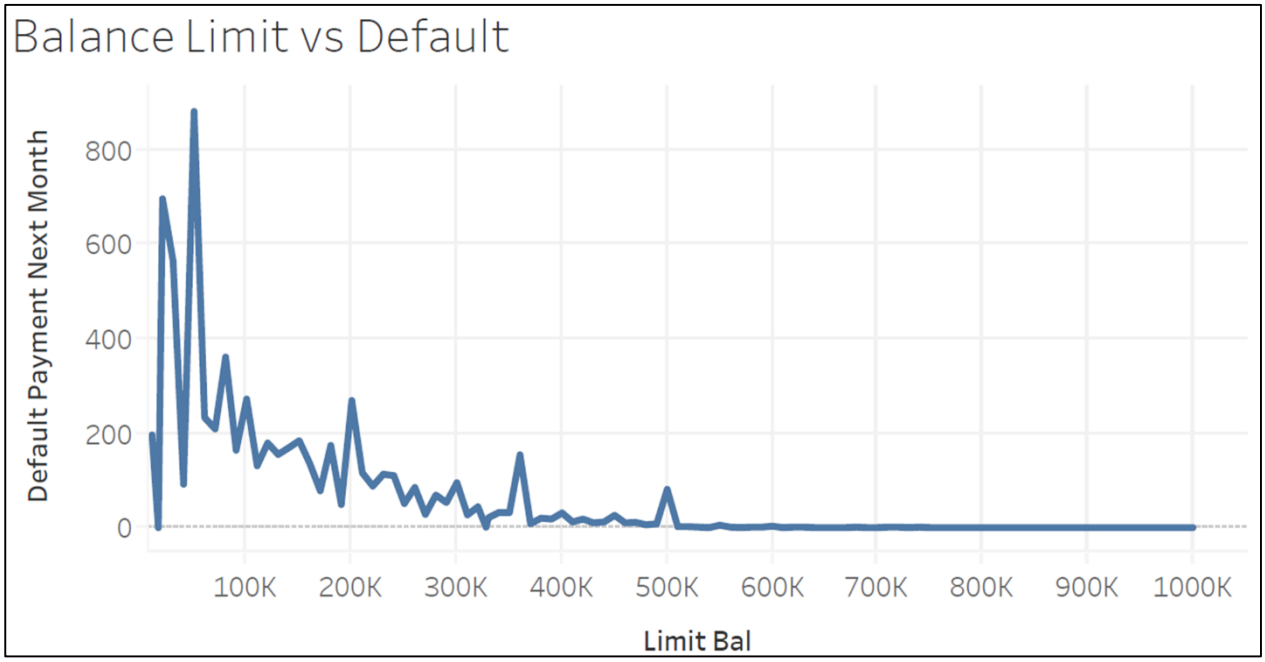




It can be seen from the visualization that the highest defaulters are customers with undergraduate degree. In the next visualization, the insight that we gather is that customers that are single default the most. However, the difference between single and married customers is not drastic.



From the above visualization, we gather that most defaulter fall in the age range 25 to 30 with a max at age 29 and a secondary peak at age 36.



From the above visualization we gather that most of the defaulters had a balance limit of below 100,000 (New Taiwan Dollars), out of which, most defaulters had a balance limit of 50,000 (New Taiwan Dollars) and a secondary peak was at 20,000 (New Taiwan Dollars).

Modelling

We used three predictive models for our project which includes one linear model, a non-linear model and a tree based model. Our problem was a classification problem and hence models best suited for the classification were chosen. They are as follows:

a. Logistic Regression with backward selection

b. K- nearest neighbor model

c. Gradient boosting trees.

Data Separation

The data set was split into training and testing data set. The model will be trained with the training data set and the model performance will be tested with the testing data set. The statistics derived from the training data set is used for model comparison.

Training: 2070 rows

Testing : 8881 rows

This same data separation was used for all the models discussed in this report.

Logistic Regression with Backward Selection

A simple logistic regression with backward selection was performed. The backward selection process was done in order to select the best combination of predictors to yield the best accuracy.

```
Call:
glm(formula = default.payment.next.month ~ LIMIT_BAL + SEX +
    EDUCATION + MARRIAGE + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_5 +
    BILL_AMT1 + BILL_AMT2 + BILL_AMT5 + PAY_AMT1 + PAY_AMT2 +
    PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6, family = binomial,
    data = training)
```

In the above picture, we can see that the backward selection process has given us the best combination of predictors that need to be considered for the highest accuracy. Only 18 of the available 23 predictors were selected. The model was then trained with the above selected predictors. As we mentioned earlier after the model was trained, it was then tested on the testing set which is about 30% of the data. The result can be observed with the confusion matrix shown below.

Confusion Matrix

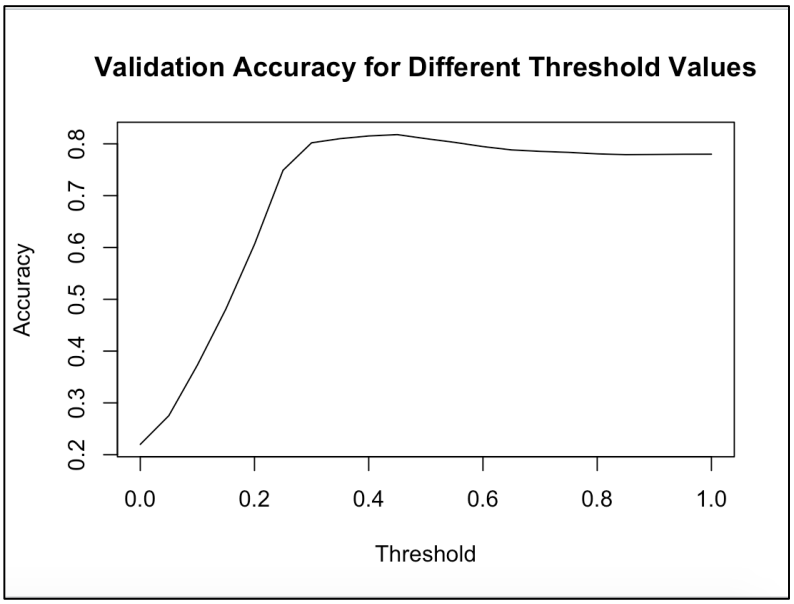
Predicted/ Observed	Non- Default	Default
Non- Default	6725	1484
Default	203	468

Here we can observe that the True Positives are 6726 instances and the True negatives are 468 instances. The following measures can be summarized from the model we just created. The accuracy we achieved is **"81%"**

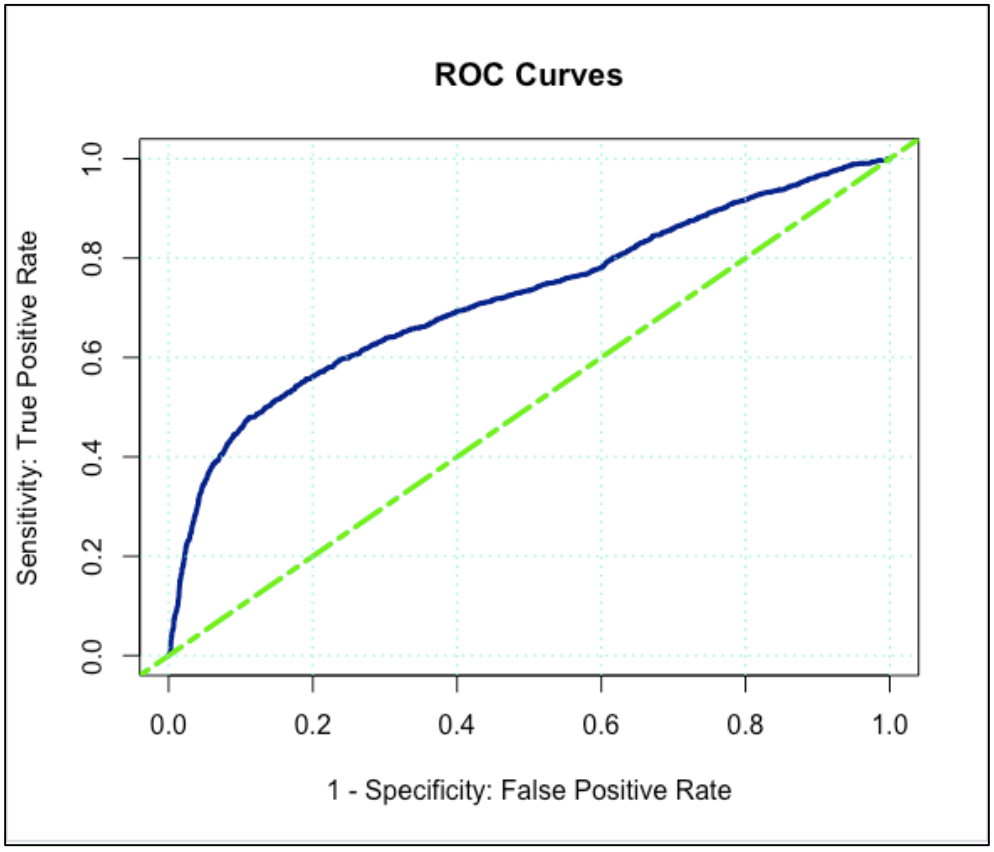
Classification Statistic:

Statistic	Logistics
Sensitivity	**0.971**
Specificity	**0.240**
NPV	**0.697**
PPV	**0.819**
AUC	**0.719**

This is a graph between the accuracy of the validation set or the testing set with respect to various threshold values. From the graph, we chose threshold as **0.5** and we can get the highest accuracy at that point.



Let us now look at the ROC curve and the AUC value

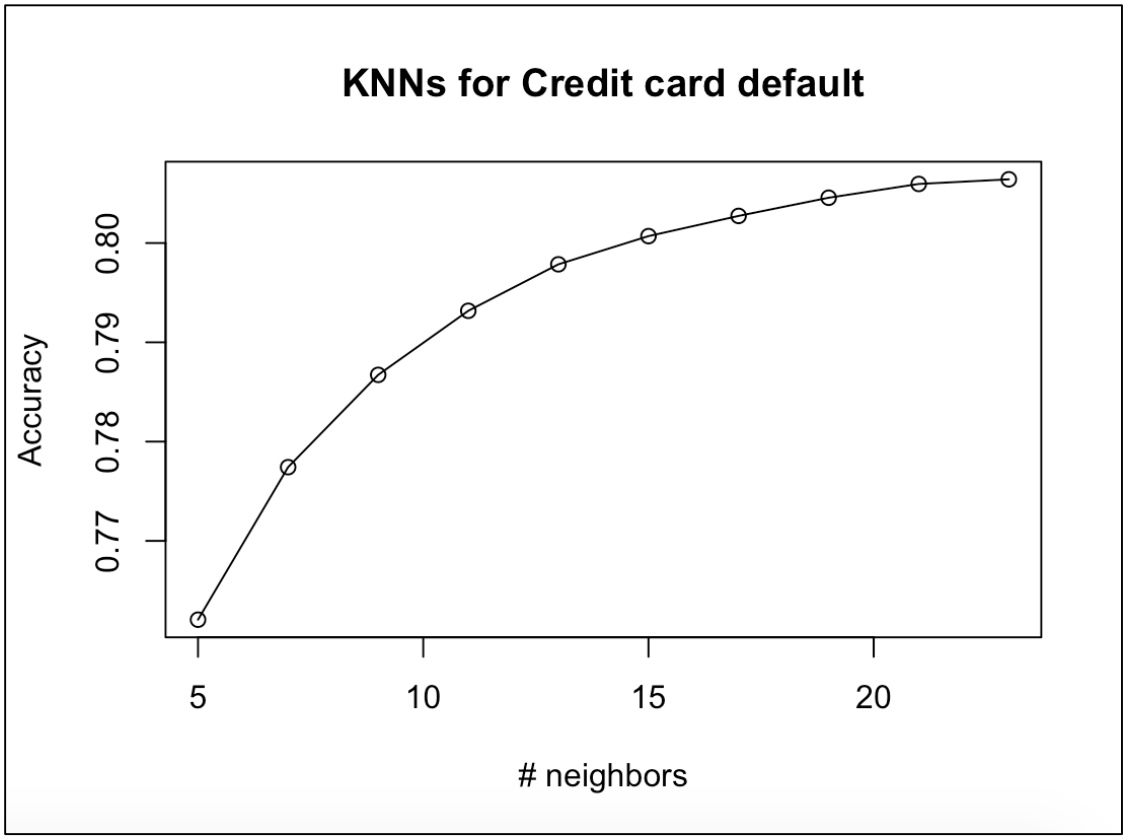


The ROC curve shows that the prediction accuracy is better than random classification. The Area under the curve is 0.71933 which is closer to one and greater than 0.5.

K-Nearest Neighbor

For the KNN-model we created a training and validation set with the results we obtained from the variable importance of Random Forest modelling technique. We went with the top 19 predictors of the variable importance plot. For the preprocessing, we had performed center and scaling method for the numerical stability. Also, the model was tested with different values of "K" and finally the "K" with best accuracy was chosen for our model.

Let us now look at the plot of accuracy measures with varying "K" values



When K = 23 we get the highest accuracy. Therefore, the system chose K = 23 and then trained the model with it. It was then tested on the validation set and the result can be seen with the help of the confusion matrix below.

Confusion Matrix

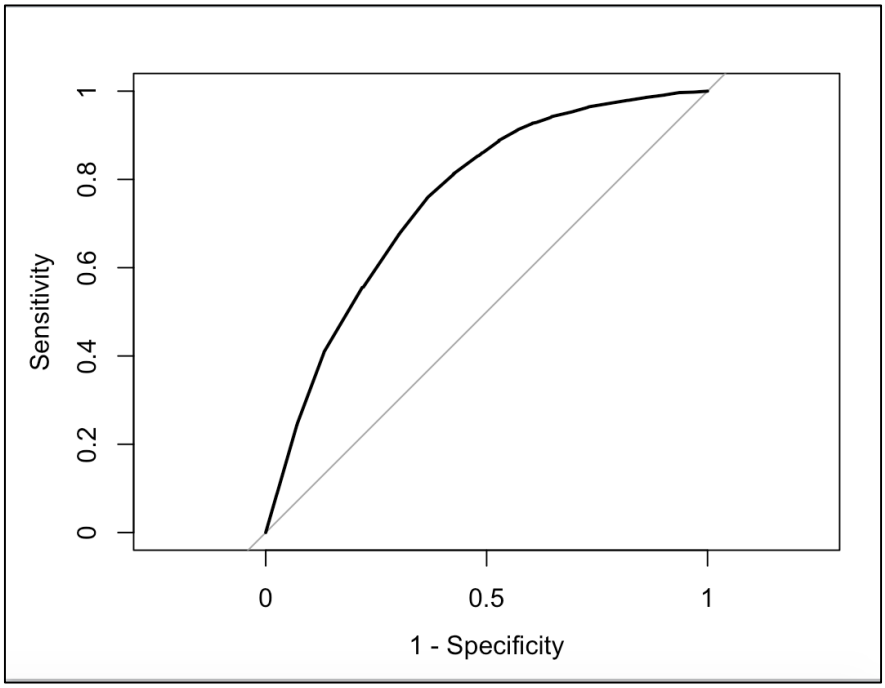
Predicted/ Observed	Non- Default	Default
Non- Default	6518	1264
Default	410	688

Here we can observe that the True positives are 6518 instances and the True negatives are 688 instances. The following measures can be summarized from the model we just created. The accuracy achieved with this model was **81.1%** which is **0.1%** better than the logistic regression model.

Classification Statistic:

Statistic	Logistics
Sensitivity	**0.941**
Specificity	**0.352**
NPV	**0.626**
PPV	**0.837**
AUC	**0.756**

Let us now look at ROC curve and the AUC value



The ROC curve shows that the prediction accuracy is better than random guess. The Area under the curve is 0.756, which is closer to one and greater than 0.5.

Gradient Boosting Trees

Similar to the KNN model we divided with training and testing set. The predictors used were from the variable importance plot of Random Forest technique and from inference gained through data exploration, which were 19 predictors. For the preprocessing, we had performed center and scaling method for the numerical stability.

The model was then trained with the above selected predictors. As we mentioned earlier after the model was trained, it was then tested on the testing data set. The result can be observed with the confusion matrix shown below.

Confusion Matrix

Predicted/ Observed	Non- Default	Default
Non- Default	6574	1242
Default	354	710

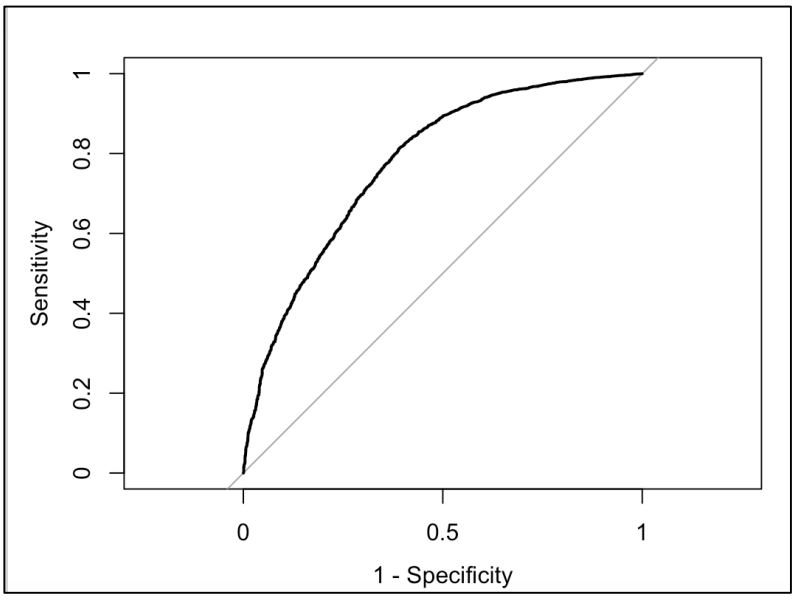
Here we can observe that the True positives are 6574 instances and the True negatives are 710 instances. The following measures can be summarized from the model we just created

Classification Statistic:

Statistic	Logistics
Sensitivity	**0.949**
Specificity	**0.364**
NPV	**0.667**
PPV	**0.841**
AUC	**0.776**

The accuracy obtained by this modelling technique is 82% which is better than both KNN and logistic regression models.

Let us know look at the ROC curve and the AUC value



The ROC curve shows that the prediction accuracy is better than random classification. The Area under the curve is 0.71933 which is closer to one and greater than 0.5.

Selecting the Important Predictors

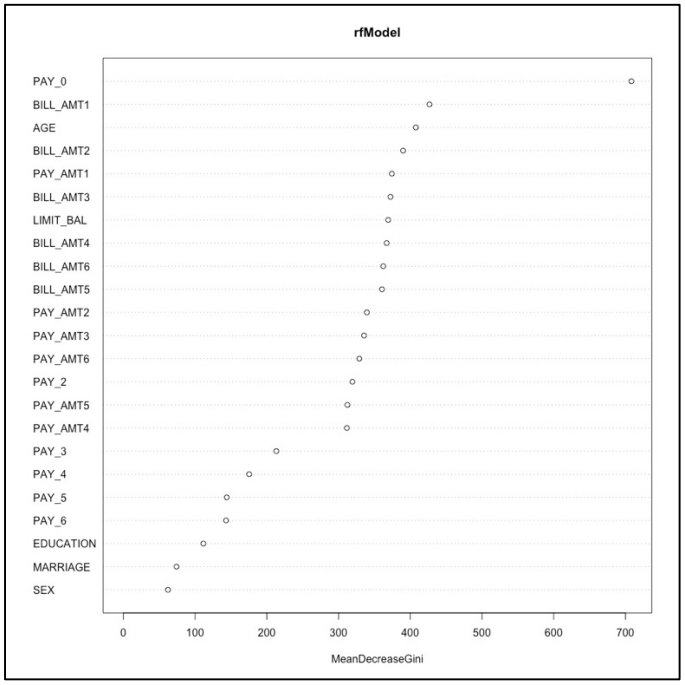
Picking the predictors which are important in predicting a defaulting customer is of prime importance because they provides focus areas for financial product designers to improve existing products and create new avenues for increased revenue with balanced risk. Selecting the best predictor set out of this exercise has two important factors, they are:

- It has to be smallest best subset available
- This best subset should give highest accuracy comparatively

The best smallest subset according to backward elimination that gives best classification statistics are

- Limit Balance
- Sex
- Education
- Marriage
- Age
- Pay 0 – Pay5
- Bill_Amt 1- 2, Bill_Amt 5
- PAY_AMT1-PAY_AMT6

Although the subset is reduced from 23 to 14 predictors, they were performing well for liner models and were not showing similar performance for other models. We further deployed methods like variable importance plot and visualization techniques to understand the predictor importance and inter predictor relation. Let us analyze the variable importance plot that was identified by a random forest model.



It is observed from the above figure that predictors like Education, Marriage and Sex are not contributing much to the model accuracy. The data visualizations made to explore the inter predictor relations have also shown that the factor variables like Education, Sex and Marriage do not explain any specific trend. By

collaborating the outputs of the above two methods and our observations form data explorations, we have reached at a conclusion that the smallest best subset is as follows:

- PAY_0-PAY_6
- BILL_AMT_1-BILL_AMT_6
- PAY_AMT_1-PAY_AMT_6

To be noted, the subset chosen were financial records of last six months.

Conclusion

This subset has 19 predictors which showcase the highest accuracy when they were used to train models. On analysis by a subject expert, it was observed that these predictors do govern the decision to grant a credit card to a new applicant. The predictors selected, provides the data about the new applicant's financial records. If the applicant has been paying his dues correctly without fail and has paid amounts close to his balance due, he/she is likely to be granted a credit card. Factors like Gender and Education are not providing enough information but the decision to remove Age was not encouraged by the human expert, but on the other hand, including the same to the subset is not improving the classification statistics. It is to be noted that by removing factor variables like Gender and Education, we are also reducing the chances of data entry error. The data analysis says that, if you have a good financial discipline you will be granted a new credit card. If not, irrespective of your educational qualification it will be risky to grant the same. As the financial records are observed, it is noted that the nature of his/her payment of dues is important and also the amount he/she pays each month. If the amount he/she paid and the amount due is positively correlated, then the applicant has a high chance of getting his/her application approved (since he/she displayed solid history of repayment capability). Apart from this, the older a financial record becomes, the lesser it is of importance in predicting default and vice versa.

Let us Compare Model Accuracies,

Model	Accuracy	Kappa	AUC
Logistics	81.0023	0.275344	0.719337
KNN	81.1486	0.347947	0.7575
Gradient Boosting	82.027	0.373681	0.7763

For our data set gradient boosting has the best accuracy statistically but for easier implementation we will be selecting logistic regression, as the difference in accuracy is not huge and logistic regression takes much less computational time. If the model is demanding higher accuracy, then Gradient Boosting is the best option. KNN is the least favored option since it takes more time to train and its model consumes more computational memory space.

Instructions

The Rcode file name is Final.R and the dataset name is Database. Please set the working directory of R studio to the location where the dataset and the Rcode is saved. Run the Code in the usual manner and the corresponding results will be generated as shown in the report.