Credit Standing Analysis

Anjukutty Joseph : R00171244

05 December 2018

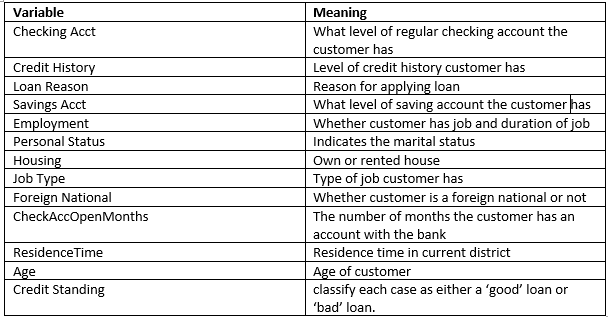
### Objective

The purpose of this analysys is to assess the credit worthiness of future potential customers in a finacial institution .

**a. Exploratory Data Analysis**  
The data set analysed in this report consists of 780 past loan customer cases, with each case classified as either good or bad loan. A scoring data set of 13 observations are used to predict the behaviour of the model. A first glance of the data set shows below details.

## 'data.frame': 780 obs. of 14 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Checking.Acct : Factor w/ 4 levels "0Balance","High",..: 4 1 1 1 4 3 1 3 1 4 ...  
## $ Credit.History : Factor w/ 5 levels "All Paid","Bank Paid",..: 1 4 4 4 1 4 1 2 1 5 ...  
## $ Loan.Reason : Factor w/ 10 levels "Business","Car New",..: 2 2 2 5 10 2 2 4 3 3 ...  
## $ Savings.Acct : Factor w/ 5 levels "High","Low","MedHigh",..: 2 2 5 5 5 4 2 2 2 2 ...  
## $ Employment : Factor w/ 7 levels "","Long","Medium",..: 3 5 2 2 2 7 2 3 2 5 ...  
## $ Personal.Status : Factor w/ 4 levels "","Divorced",..: 4 2 2 1 4 2 3 2 4 3 ...  
## $ Housing : Factor w/ 4 levels "","Other","Own",..: 3 3 3 3 2 3 3 2 1 4 ...  
## $ Job.Type : Factor w/ 4 levels "Management","Skilled",..: 1 2 2 2 2 4 2 4 2 2 ...  
## $ Foreign.National : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 2 2 ...  
## $ Months.since.Checking.Acct.opened : int 7 16 25 31 7 13 22 25 25 13 ...  
## $ Residence.Time..In.current.district.: int 3 2 2 4 4 2 3 4 4 4 ...  
## $ Age : int 44 28 28 30 35 22 29 33 62 40 ...  
## $ Credit.Standing : Factor w/ 2 levels "Bad","Good": 2 1 1 2 2 2 2 2 2 2 ...

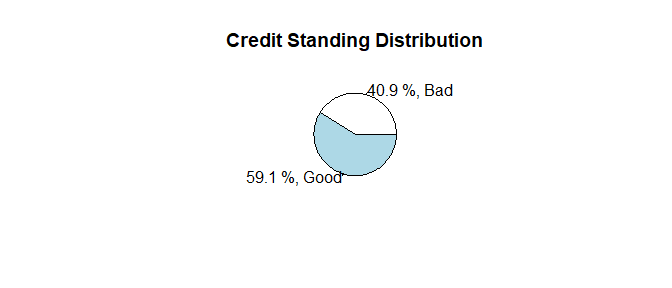
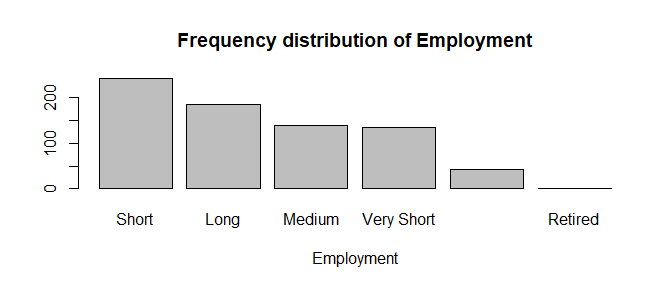
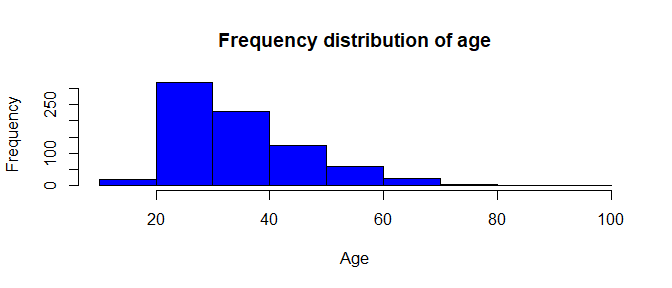
The data set contains 13 customer attributes, 10 attributes are categorical and 3 are continuous. ID column is ignored in this analysis. Categorical variables are factor type and others are numerical. Columns- ‘Employment’, ‘Personal Status’ and ‘Housing’ are found to have some missing values. Among them ‘personal status’ and ‘Housing’ have less than 1% missing and ‘Employment’ have around 4% missing values out of total records. Since the percentage of such records are few and also to keep it simple, ‘NA’ is assigned to missing values. The names and meaning of each variable are illustrated in the table below. Two variables having long names are renamed for ease of use.



### Data Visualisation

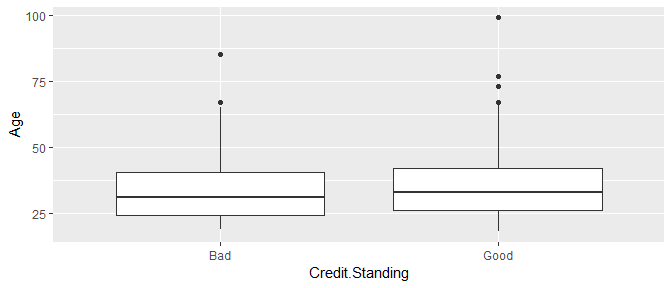
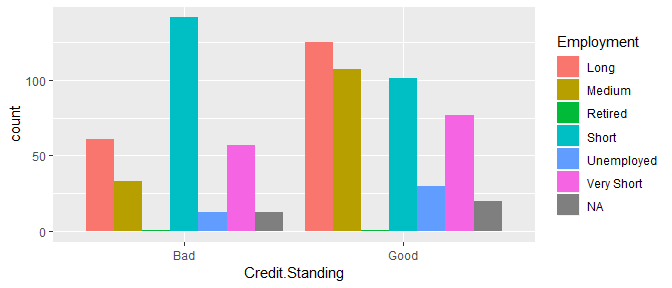
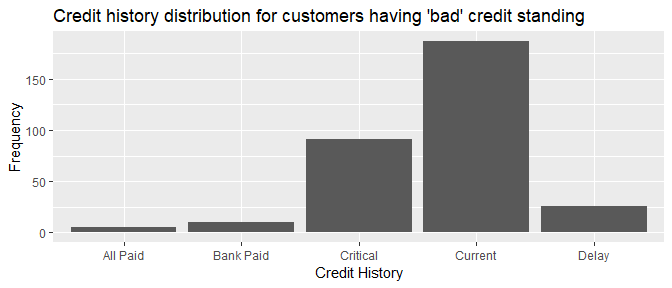
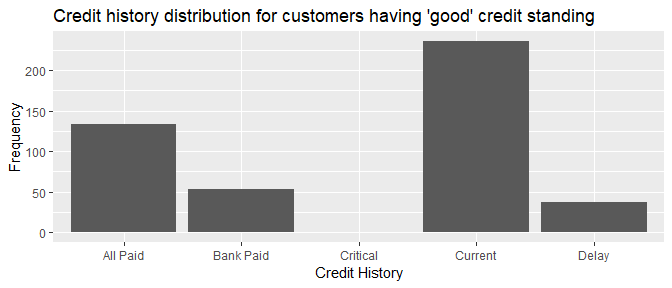
To have an understanding of various attributes and relationship among them, following visual analysis are performed.

#### Univariate Analysis

  The credit standing disrtbution shows that more loans are good than bad in the given data set. So the baseline model will be correct 59.1% of the time. The aim is to create a model which gives more accuracy than this baseline. From frequency distribution of emplyoments, the number people having short term employment are more. Age is skewed and doesn’t look normally distributed.

#### Bivariate Analysis

To understand the influence of individual variables on credit standing, below plots are made.



From the barplot of credit standing and credit history, it is clear that majority of customers in the data set having ‘critical’ credit history have ‘bad’ credit standing. Those who have ‘All Paid’ credit history are more likely to have good credit standing. So the variable ‘credit history’ has a significance in determining credit standing. A graph of employment aganist credit standing is made to check for any pattern.It is difficult to generalise but variable ‘Employment’ seems to have some significane in predicting credit standing.

From the boxplot of Age and credit standing, those who have good credit standing have slightly higher median age when compared to people having bad credit.

#### Trivariate Analysis

It is difficult to get a clear vision of the effect of other variables in the data set on credit standing. However,the variable ‘Foreign National’ seems to have some predicting capacity. So a trivariate analysis of ‘credit standing, credit history and Foreign national is performed using ’flat’ contingency tables to get a more clear sense of the relationship. Its matrix representation is represented below.

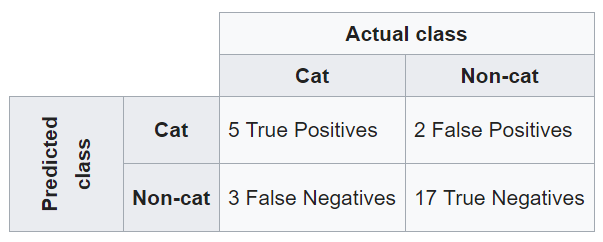
## ForeignNational No Yes  
## creditStanding CreditHistory   
## Bad All Paid 4 1  
## Bank Paid 4 6  
## Critical 13 78  
## Current 69 118  
## Delay 16 10  
## Good All Paid 29 104  
## Bank Paid 13 40  
## Critical 0 1  
## Current 87 149  
## Delay 18 20

It is interesting to see that, there are more foreign customers than local residents and only 15% foreign nationals have critical credit history and majority of them have good credit standing.

### b. ROC Curve(Receiver Operating Characteristic)

ROC curve serves to measure the perfomance of the classification model.This is used as a perfomance metric for classification problems. The x axis of ROC curve represents False positive Rate (FPR) and y axis represents True positive rate (TPR).  
**FPR** : Percentage of records that the model classifies as positive but are actually negative.  
**TPR** : Percentage of records that the model classifies as positive and are positive  
FPR and TPR can be calculated from confusion matrix. For the confusion matrix displayed below FPR and TPR are calculated.

##### Confusion Matrix



True Positive Rate(TPR) for the above confusion matrix = TP/TP+FN = 5/(5+3)  
False Positive Rate(FPR) for the aabove confusion matrix = FP/FP+TN = 2/(2+17)

A sample image of ROC curve is shown below. It shows the behavior of classification model for all possible threshold. A threshold is a value chosen to classify each record in the data set to which class it belongs to. ROC curve captures the variations in TPR and FPR with the change in threshold. For a best model, TPR should be high and FPR should be low.

Area under ROC curve is a measurement of how good the model is in classifying the two outcomes. For example, in credit standing analysis, higher AUC of ROC curve indicates the final model is better in predicting good credit standing as good and bad credit standing as bad. An AUC near to 1 shows a best model. The middle blue line represents kind of no information line and in such case AUC will be 0.5 which means the model has no class separation capacity.

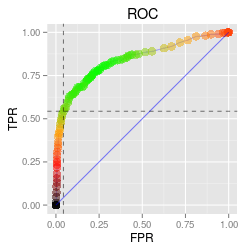
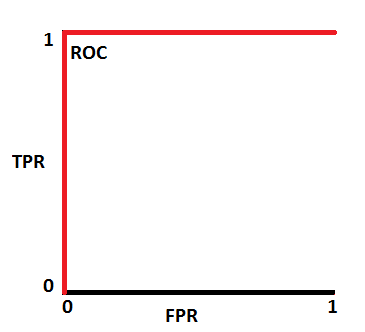
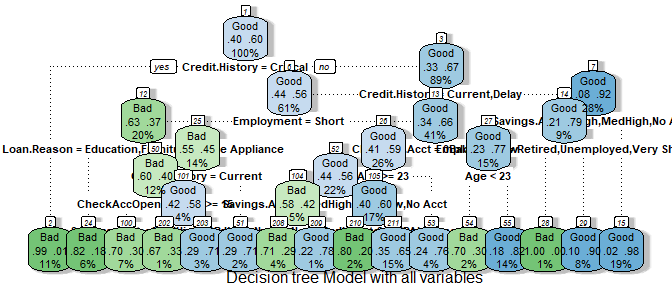
 

Fig :ROC Curve with AUC =1

## C. Decision tree

To obtain an accurate model, using supervised learning technique, a decision tree is built to decide credit standing. This model is built by considering all variables except ID to predict credit standing. The data set is split into 80:20 ratio randomly for training and test of the model. The picture of decision tree model is displayed below.



As shown in the tree diagram out of 12 predictor variables, 7 variables namely “Credit.History” ,“Employment”,“ResidenceTime”, “Loan.Reason”,“Checking account”,“age” and “savings Acct” are involved in building this tree model. These variables are considered to be most significant in predicting Credit Standng for customers by this model. This model is trained on 624 samples.There are 16 terminal nodes which classifies a given customer has good or bad credit standng. This decision tree classifies customers having ‘critical’ credit history to have bad credit standing without any further checking on other variables. A summary of the model is displayed below.

##   
## Classification tree:  
## rpart(formula = Credit.Standing ~ . - ID, data = train\_tree\_data)  
##   
## Variables actually used in tree construction:  
## [1] Age CheckAccOpenMonths Checking.Acct   
## [4] Credit.History Employment Loan.Reason   
## [7] Savings.Acct   
##   
## Root node error: 251/624 = 0.40224  
##   
## n= 624   
##   
## CP nsplit rel error xerror xstd  
## 1 0.274900 0 1.00000 1.00000 0.048801  
## 2 0.065737 1 0.72510 0.77689 0.046130  
## 3 0.013944 3 0.59363 0.59761 0.042527  
## 4 0.011952 5 0.56574 0.68526 0.044470  
## 5 0.010000 15 0.43426 0.66135 0.043977

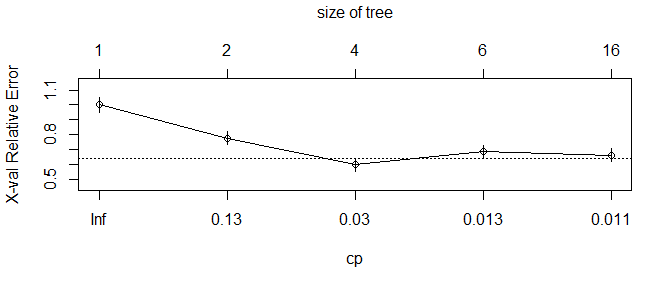
The confusion matrix displayed below shows an accuracy of 76.9% for this model when tested with test data set. This is better than baseline which was 59.1%.

##   
## tree\_Model1\_pred Bad Good  
## Bad 50 18  
## Good 18 70

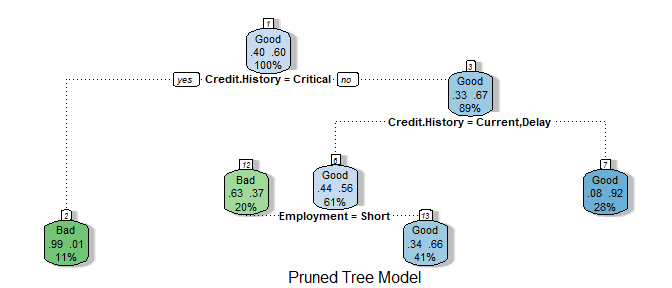
While doing exploratory data analsys in first step,the variables credit history, loan reason, foreign national and employment appeared to have significance in deciding credit standing. But, The decision tree model created with these attributes resulted in less accuracy .The misclassification rate for that model was high. So the model with selected attributes are ignored.

##### Pruning

Even though this tree model gives 76.9% correct predictions on test data set, it is a bit bushy. The misclassification of actual bad as good is 26.47% and actual good as bad is 20.45%. Inorder to reduce misclassification error, an attempt is made to prune this model by taking misclassification as tuning parameter.



From the graph above, the relative error for the decision tree model is least at cp =0.03. So, to reduce misclassification, this cp value is picked to prune the tree. The pruned tree model is displayed below.



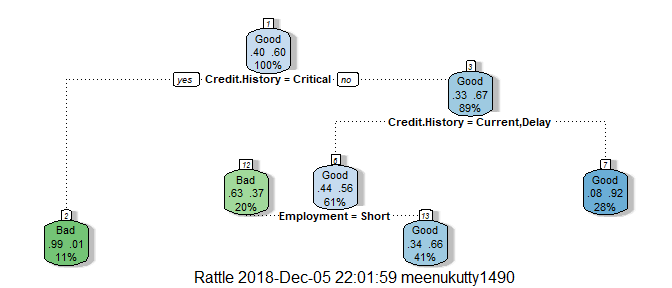
##   
## prune.tree\_Model1\_pred Bad Good  
## Bad 45 8  
## Good 23 80

The resultant pruned model has accuracy of 80% when tested with test data. This is an improvement from the first decision tree model built using all variables.The pruned model has 4 leaf nodes and decision is taken based on two variables- credit history and Employment. These two variables are considered most relevant in predicting the credit standing of customer. Another great advantage after pruning is that, the tree is more shallow and is easily readable.

### d. Test Model on Scoring data set.

In this section, the pruned tree model is used to predict credit standing potential customers in the scoring data set. The model is applied to a scoring data having 13 records each one corresponding to 13 potential clients. The working of the model in deciding credit standing is explained below by choosing 3 customers.

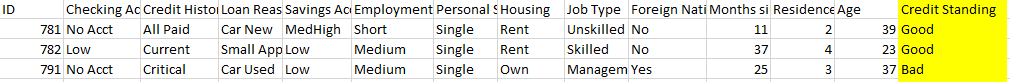
The model is in the shape of a tree as illustrated below. It essentially classifies data into smaller subsets and finally reach to a conclusion. It consists of decision nodes and leaf nodes. The leaf node represent the results which in this analysis is ‘good’ or ‘bad’ credit standing. The decision node applies a condition to a given record and partition it accordingly. The top most node represent the most significant variable or best best predictor variable which is credit history.

 The predicted credit standing for 13 clients are as below.

modelPrediction

## 1 2 3 4 5 6 7 8 9 10 11 12 13   
## Good Good Good Good Good Good Good Good Good Good Bad Bad Bad   
## Levels: Bad Good

Among them, following 3 customer’s credit standing prediction is taken into account to explain the decision making process.



Customer1: First check at root node is whether customer has critical credit history.Since this customer has credit history of “All Paid”, it is routed to right child node corresponding to ‘no critical credit history’. The next decision node checks whether this customer has either ‘current’ or ‘Delay’ credit history. The decision at this node is ‘no’ since it is ‘All Paid’ and hence this customer is routed to its right child node which is the result leaf node the tree classify the credit standing of customer as ‘good’. The prediction at this leaf node was correct 92% of the time,while training this model which is indicated in the box.

Customer2. This customer has ‘Current’ credit history. So the first check yields result ‘no’ so tree inturn takes it to right child node of root node. This node checks if customer is having ‘current’ or ‘Delay’ credit history. This condition will give result ‘yes’ and the model takes the customer to left child node of current node. This node consider another attribute of customer which is ‘short term’ employment. Since this customer has ‘medium’ employment the model predicts ‘good’ credit standing.The prediction at this leaf node was correct 66% of the time,while training this model which is indicated in the box.

Customer 3: The most significant deciding factor according to this model is ‘critical’ credit history. As this customer has’ critical’ history the top most node check gives ‘yes’ and tree model route the customer to left child node which predicts the credit standing is going to be bad. 99% of the time, the predictions was correct at this node while training this model.

#### Accuracy Calculation

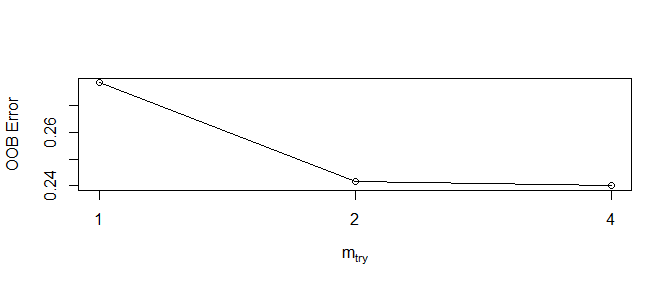
The accuracy of the model is calculated by testing it using test data set. This test data set was not part of model building. From the confusion matrix below , the model was capable of predicting correct result 80% of the time. Confusion matrix shows actual values aganist predicted values by model. (45+80)/156 = 80.1% .

##   
## Bad Good  
## Bad 45 8  
## Good 23 80

## E. Ensemble techniques

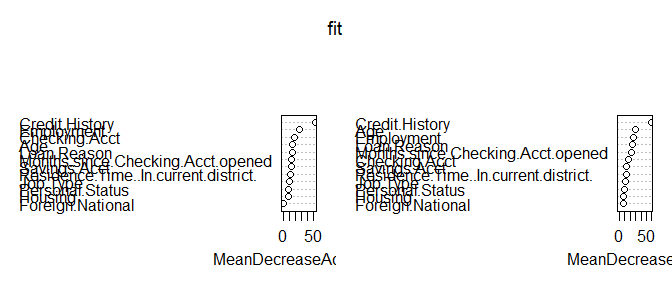
An improvement in model is tried using random forest ensemble technique. Random forest fits a number of trees on various sub samples of the data set and averages the output to increase accuracy of result. A random sample of 80% is chosen to train the model. As depicted in the graph, OOB error is least when mtry =2. So when doing a split 2 variables are chosen random which best suits the split by this model. Number of trees is 500.

## mtry = 1 OOB error = 27.83%   
## Searching left ...  
## Searching right ...  
## mtry = 2 OOB error = 24.17%   
## 0.1317365 0.05   
## mtry = 4 OOB error = 24%   
## 0.006896552 0.05



## mtry OOBError  
## 1.OOB 1 0.2783333  
## 2.OOB 2 0.2416667  
## 4.OOB 4 0.2400000

##   
## fitPrediction Bad Good  
## Bad 40 12  
## Good 29 56

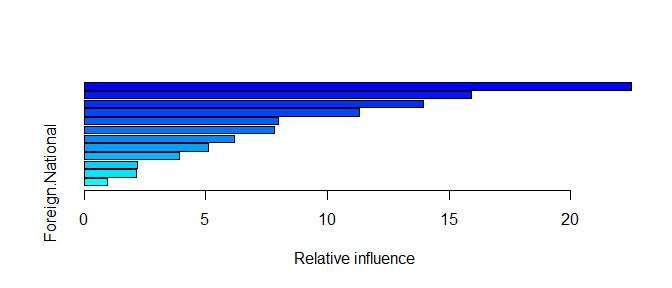


From variable importance plot , credit history has most significance and Foreign national has least. The accuracy of random forest model when tested on validation data set is 70%. This is less than accuracy of pruned decision tree model. So, decision tree model is best as of now. The tuning parameters provided to build the model are, ‘mtry’ =2 as it shows least OOB error and number of trees as 5000. The poor performance of tree can be due to the dataset analyzed here. In addition to that, this model did not consider any other tuning parameters like ‘node size’, ‘classwt’ etc.

## Boosting

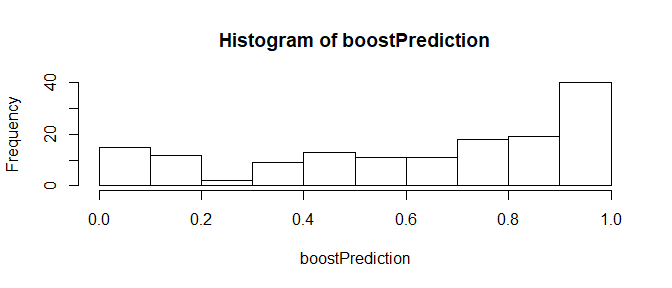
Another ensemble technique ‘boosting’ is tried to build a suitable and accurate model from data set. Boosting is a sequential execution in which each model is generated by trying to reduce the bias created in the previous model.

Random sample of 80% of training data is taken to train the model using gradient boosting. The accuracy of this model is 70% when applied on test data. Although boosting is a very powerful technique, it doesn’t seem to help here in improving the accuracy. This can be because of the variance in data set as it was only 630 records. Another reason could be selection of inappropriate tuning parameters so that, it could not generalize so well to convert a weak learner to strong. This model was built by supplying tuning parameters as following. number of trees 5000, shrinkage= 0.01 and interaction depth which is the number of splits as 2.



**Variable Importance Plot**

## var rel.inf  
## Credit.History Credit.History 22.5093885  
## Employment Employment 15.9333125  
## Loan.Reason Loan.Reason 13.9343396  
## Age Age 11.3253162  
## Personal.Status Personal.Status 7.9908207  
## CheckAccOpenMonths CheckAccOpenMonths 7.8120398  
## Savings.Acct Savings.Acct 6.1788420  
## Checking.Acct Checking.Acct 5.0903273  
## ResidenceTime ResidenceTime 3.9303394  
## Housing Housing 2.1800069  
## Job.Type Job.Type 2.1538010  
## Foreign.National Foreign.National 0.9614661



**Confusion Matrix of Boosting Model**

##   
## predict\_class Bad Good  
## bad 37 14  
## good 29 70

## 

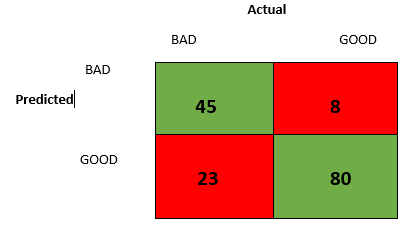
From the ROC curve and Confusion matrix, it is visible that the boosting model misclassified many actual bad ones as good ones. So, the FPR is high.

## f. Misclassification.

Classification Error: In the given data set, around 60% of customers have good credit standing and around 40 % have bad. This is almost fairly balanced. But, there are chances that there is big gap between good class and bad class. For example, if 95% of customers in the analyzed data set was good and only 5% have bad credit the model accuracy will be misleading.

Classification Error = FP/(FP+TN) + FN/(FN+TP)

The confusion matrix for the best model(decision tree),is as shown below.The red boxes indicates misclassified data.



Given, cost of misclassifying actual bad as good is :1 cost of misclassifying actual good as bad is : 5

Using the confusion matrix above , the best model classified 23 out of 68 bad customers as good. Also, it classified 8 out of 88 good customers as bad. So the average cost per classified record is (23*1+8*5)/156 = 0.4 To understand more on misclassification, Linear Discriminant Analysis and quadratic discriminant analysis is performed. Summary is displayed below.

**LDA**

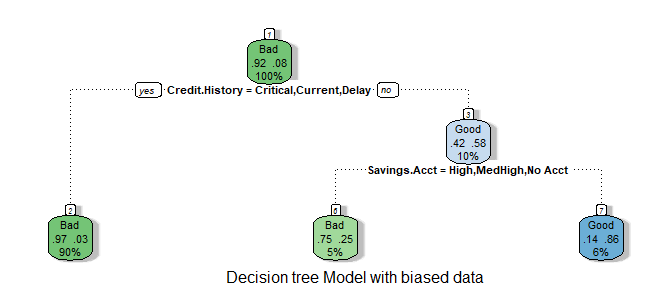
## Call:  
## lda(Credit.Standing ~ Credit.History + Employment, data = RFtrain\_tree\_data)  
##   
## Prior probabilities of groups:  
## Bad Good   
## 0.3933333 0.6066667   
##   
## Group means:  
## Credit.HistoryBank Paid Credit.HistoryCritical Credit.HistoryCurrent  
## Bad 0.02966102 0.292372881 0.5720339  
## Good 0.11813187 0.002747253 0.5082418  
## Credit.HistoryDelay EmploymentMedium EmploymentRetired  
## Bad 0.08898305 0.1101695 0.000000000  
## Good 0.07692308 0.2472527 0.002747253  
## EmploymentShort EmploymentUnemployed EmploymentVery Short  
## Bad 0.4491525 0.04661017 0.1864407  
## Good 0.2225275 0.07142857 0.1923077  
##   
## Coefficients of linear discriminants:  
## LD1  
## Credit.HistoryBank Paid -0.51187933  
## Credit.HistoryCritical -3.95323391  
## Credit.HistoryCurrent -1.60013569  
## Credit.HistoryDelay -1.72403728  
## EmploymentMedium 0.34532636  
## EmploymentRetired 0.52655622  
## EmploymentShort -0.78164633  
## EmploymentUnemployed 0.01112861  
## EmploymentVery Short -0.08414517

##   
## Bad Good  
## Bad 42 12  
## Good 27 56

SUMMARY of QDA Model

## Call:  
## qda(Credit.Standing ~ Credit.History + Employment, data = train\_tree\_data)  
##   
## Prior probabilities of groups:  
## Bad Good   
## 0.40301 0.59699   
##   
## Group means:  
## Credit.HistoryBank Paid Credit.HistoryCritical Credit.HistoryCurrent  
## Bad 0.04149378 0.26970954 0.6016598  
## Good 0.13445378 0.00280112 0.4929972  
## Credit.HistoryDelay EmploymentMedium EmploymentRetired  
## Bad 0.07053942 0.1120332 0.004149378  
## Good 0.07563025 0.2352941 0.002801120  
## EmploymentShort EmploymentUnemployed EmploymentVery Short  
## Bad 0.4605809 0.04149378 0.1742739  
## Good 0.2408964 0.06162465 0.1820728

To investigate the effect of misclassification, a biased dataset is created where 94% of records have ‘bad’ credit standing and 6% have ‘good’ credit standing. The new model is shown below.

 The confusion matrix gives the accuracy of the model as 96.7%. But the high accuracy is not trustable because model is trained on a biased data. it classified 60% of good customers as bad. When such a model is applied to records which has more probable good customers, there are more chances that the model classifies them as bad.

##   
## bTreeModel\_pred Bad Good  
## Bad 89 2  
## Good 1 1

Since the cost of losing a good customer is more than the cost of giving loan to bad customer, this model is not good in predicting credit standing.

## g. Incorrect Pattern

In this section, an effort is made to find series of consecutive ID numbers, where the data set shows an incorrect pattern for credit Standing. To find suspicious pattern, I took the decision tree model which was trained on random 80% of the given data set. Then this model is applied to each subset of 10 records taken sequentially from the data set to predict the credit standing. My assumption is that, if there was a series of consecutive records of 10 or more where the grading in the given data set show a suspicious incorrect pattern, then the accuracy for preciction on that subset of 10 by the model will be quite low or approximately 0.

To implement this thought, I wrote the function testPattern() displayed below. It creates confusion matrix on each sample of 10 records and print the accuracy of prediction of the model. Upon analysing the accuracy, the records having ID numbers between 301- 316 and 711-721 showed lowest accuracy among all. So to conclude, the outcome of these records are incorrectly recorded in the given data set. I would have tried to remove these records from the data set and rebuild the model. I guess the overall accuracy of the model would have improved if I do so.

library(caret)

## Loading required package: lattice

library(e1071)  
testPattern <- function()  
{  
for( i in seq(1,780,10))  
{  
 # take sample of 10 consecutive ID numbers  
 sample = subset(CreditReport, ID>i & ID< i+10)  
 patternPrediction = predict(tree\_ModRpart, sample, type = "class")  
 #Create confusion matrix and print the accuracy  
 cM = confusionMatrix(table(patternPrediction, sample$Credit.Standing))  
 #print(c(i,cM$overall['Accuracy'])) # This commented so that report looks good.  
}  
}  
testPattern()

## Conclusion

From the analysis of given data set, the most important predictor variables to decide credit standing are, credit history and employment of the customers. This observation was consistent among all models created. Pruned classification tree model gave the maximum accuracy out of all models.

## References

1. <http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html>
2. <https://www.researchgate.net/publication/321002603_Credit_Approval_Analysis_using_R>
3. [https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 4](https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5%204)
4. <https://en.wikipedia.org/wiki/Confusion_matrix>
5. <https://www.youtube.com/watch?v=X4VDZDp2vqw>
6. <http://scg.sdsu.edu/logit_r/>
7. <https://stats.stackexchange.com/questions/193424/is-decision-tree-output-a-prediction-or-class-probabilities>
8. <https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/>