DS Report-2

Vaishnavi Kale

03/12/2019

# Introduction

The data is about the financial institution.It has a total of 780 past loan customer cases each having 14 features.The attributes are financial standing, reason for the loan, employment, demographic information, foreign national, years of residence in the district and the outcome variable Credit Standing which has two categories a good loan and bad loan. In this we will build decision tree model for binary classification of outcome variable and predict the credit standing for ten new customers using the decision tree.Furthermore, we will try to improve the tree model by using random forest and boosting technquie adaboost.

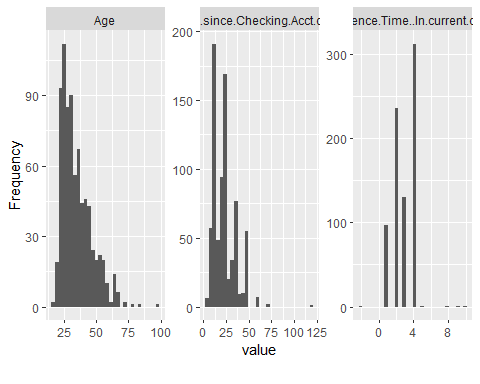
# Exploratory Data Analysis:

Let do some explorartory data analysis on the bank data.Before plotting we will see the strcture of data.It is given below:

## 'data.frame': 780 obs. of 14 variables:  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Checking Acct : chr "No Acct" "0Balance" "0Balance" "0Balance" ...  
## $ Credit History : chr "All Paid" "Current" "Current" "Current" ...  
## $ Loan Reason : chr "Car New" "Car New" "Car New" "Furniture" ...  
## $ Savings Acct : chr "Low" "Low" "No Acct" "No Acct" ...  
## $ Employment : chr "Medium" "Short" "Long" "Long" ...  
## $ Personal Status : chr "Single" "Divorced" "Divorced" NA ...  
## $ Housing : chr "Own" "Own" "Own" "Own" ...  
## $ Job Type : chr "Management" "Skilled" "Skilled" "Skilled" ...  
## $ Foreign National : chr "No" "No" "No" "No" ...  
## $ Months since Checking Acct opened : num 7 16 25 31 7 13 22 25 25 13 ...  
## $ Residence Time (In current district): num 3 2 2 4 4 2 3 4 4 4 ...  
## $ Age : num 44 28 28 30 35 22 29 33 62 40 ...  
## $ Credit Standing : chr "Good" "Bad" "Bad" "Good" ...

Age,Residence time and months since account is opened are numberical varibles.Rest all are categorical varibales which can be converted into factors before building tree model.

First we will plot some histograms to check data distribution of the numeric variables.Here, there are three histograms of age,months since checking account opened and Residence time.



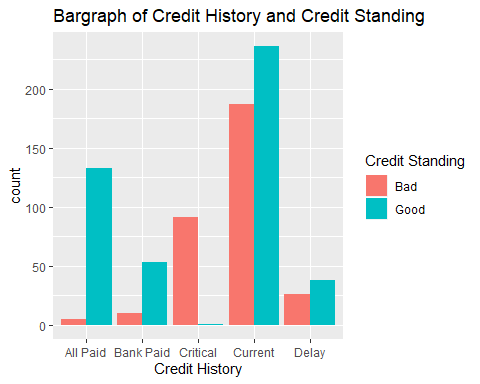
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 26.00 32.00 34.75 41.00 99.00

From first graph which is of age variable shows that the data distribution is right skewed. This can be also be seen by summary:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 26.00 32.00 34.75 41.00 99.00

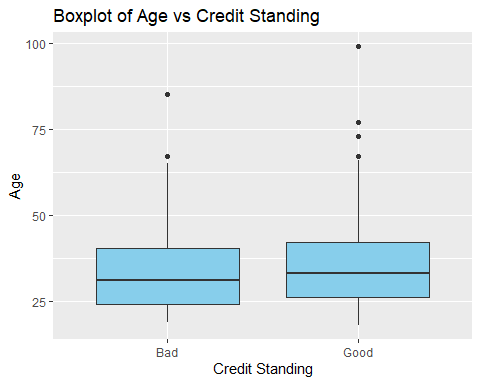
The centrality of the data is given by the mean which is 32. Most of the age data is between 25 and 50 years. Second graph is histogram of Months since checking account opened,here most of the data lies from 0 to 50 months and the data distribution is right skewed. Third graph illustrate that maximum number of people have four years of residence time.Followed by two years of residence and there are quite few people with more than 4 years of residence time.

To check the relation between credit history and credit standing we will plot a bargraph.It is as follow:



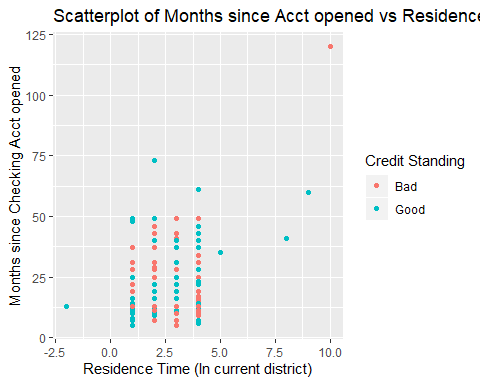
From above graph, it is clear that a large number of people with credit standing either good or bad have credit History as current.People with credit history as critical mainly have bad credit standing.If the loan is all paid, there is high propability of getting credit standing as good.Credit history as all paid,bank paid,current and delay has high number of people with good loan.

To examine the variables credit standing and age, boxplot is given below:



Above graph depicts the relation between age and credit standing.Both the credit standing levels have almost same interquartile and actual range.Outliers are present in the data hence median and IQR gives the best measure of the centrality and variability respectively.There is very less difference between medians of bad and good.

To check the relationship between three variables Months since Checking Acct opened,Residence Time (In current district) and Credit Standing we will use scatterplot.



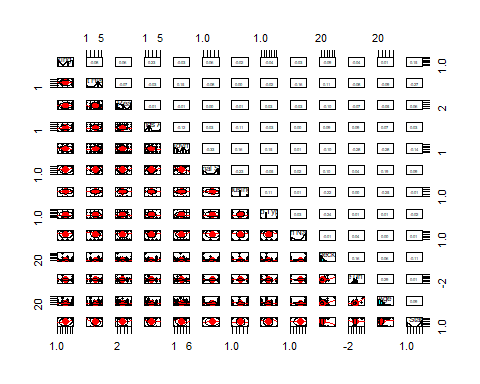
From above graph, there is no significant relation between Months since Checking Acct opened,Residence Time (In current district) and Credit Standing. Most of the outliers here have good credit standing.

To do binary analysis, below is table of employment and credit standing.

## Employment  
## Credit\_Standing Long Medium Retired Short Unemployed Very Short  
## Bad 61 33 1 141 13 57  
## Good 125 107 1 101 30 77

The table above shows that a notably large amount of people have a short employment. People with long employment mostly have good credit standing followed by medium employment while short employment people mainly have bad credit standing.

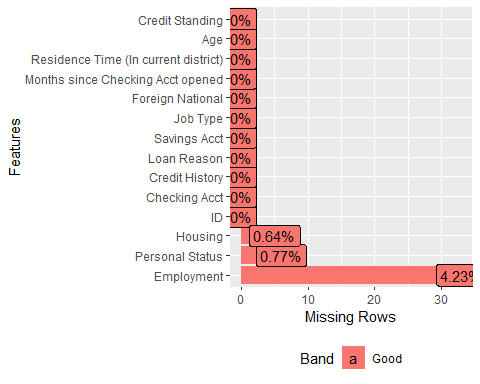
The below graph shows correlation between different variables with their distribution:



Most of the variables does not have a correlation or have a neglibile correlation. Only few are significant such as personal status and employment have a negative correlation whereas age and residence time have negative correlation.

## Data Cleaning:

There are some missing values in the data we need to deal with missing values before building our decision tree model.For this missing values are plotted below:



Personal status,Employment and Housing have some missing values. Employment have 4.23% missing values.As personal status and housing have just 6 missing values we will use statistical approch to handle these missing values.In this approach, the NA are filled with most frequent value also called mode of the data for a categorical variable.Divorced has highest frequncy in personal status hence all NA of that column are filled with it.Similary, own is filled in for all housing missing values.

As employment have high missing values, kNN imputation can used for predicting missing values of employment.kNN is k-nearest neighbors algorithm.Here, employment is taken as outcome variable with all other variables are taken as predictors and the model is constructed.This model is then used for predicting missing values of employment.

Summary of all variables :

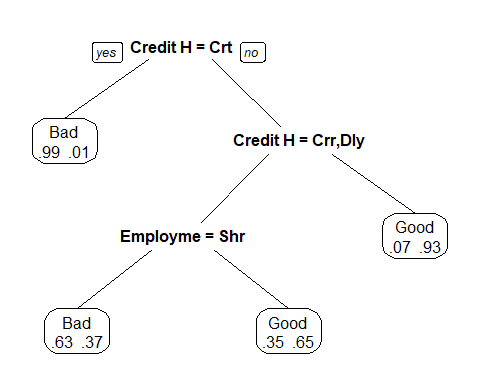
## ID Checking Acct Credit History Loan Reason   
## Min. : 1.0 0Balance:263 All Paid :138 Car New :191   
## 1st Qu.:195.8 High : 45 Bank Paid: 63 Small Appliance:186   
## Median :390.5 Low :201 Critical : 92 Furniture :160   
## Mean :390.5 No Acct :271 Current :423 Business : 82   
## 3rd Qu.:585.2 Delay : 64 Car Used : 75   
## Max. :780.0 Education : 45   
## (Other) : 41   
## Savings Acct Employment Personal Status Housing   
## High : 29 Long :195 Divorced:279 Other: 94   
## Low :500 Medium :147 Married : 70 Own :529   
## MedHigh: 49 Retired : 2 Single :431 Rent :157   
## MedLow : 80 Short :257   
## No Acct:122 Unemployed: 43   
## Very Short:136   
##   
## Job Type Foreign National Months since Checking Acct opened  
## Management:101 No :253 Min. : 5.0   
## Skilled :498 Yes:527 1st Qu.: 13.0   
## Unemployed: 19 Median : 19.0   
## Unskilled :162 Mean : 23.2   
## 3rd Qu.: 29.5   
## Max. :120.0   
##   
## Residence Time (In current district) Age Credit Standing  
## Min. :-2.000 Min. :18.00 Bad :319   
## 1st Qu.: 2.000 1st Qu.:26.00 Good:461   
## Median : 3.000 Median :32.00   
## Mean : 2.868 Mean :34.75   
## 3rd Qu.: 4.000 3rd Qu.:41.00   
## Max. :10.000 Max. :99.00   
##

Here, residence time has mininum value of -2 which is not possible since time can never be negative value. This must have been mistakely taken so we will replace it by the mean value of residence time variable.

# Desicion tree

Let start with our tree construction, the bank data is divided into two set named as training set and test test. 80% training data and 20% testing data. Train dataset is used for building decision tree model and testing set is use for doing predictions by the tree model. Credit history is our outcome or class variable and all other variables are predictors.

## n= 622   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 622 256 Good (0.41157556 0.58842444)   
## 2) Credit History=Critical 76 1 Bad (0.98684211 0.01315789) \*  
## 3) Credit History=All Paid,Bank Paid,Current,Delay 546 181 Good (0.33150183 0.66849817)   
## 6) Credit History=Current,Delay 381 170 Good (0.44619423 0.55380577)   
## 12) Employment=Short 130 48 Bad (0.63076923 0.36923077) \*  
## 13) Employment=Long,Medium,Unemployed,Very Short 251 88 Good (0.35059761 0.64940239) \*  
## 7) Credit History=All Paid,Bank Paid 165 11 Good (0.06666667 0.93333333) \*



From the above result and decision tree we can say that Credit History is root of the decision tree as it has the lowest entropy among all other variables.Only two variables are used for tree construction and has level as 4. (Prunung of tree is not required here). Bad and good are terminal nodes. Looking at tree,If credit history is critical then credit standing is bad no further conditions checked as it is leaf node and .99 indicates 99% of records in the dataset with critical history has credit standing = bad.Otherwise, check if it is current or delay, if no then credit standing=Good (This is true for 93% of the data),else go to employment node.here check whether employment is short for yes again check credit history else it will come to leaf node Good.This is how decision tree works. We will use this decision tree to predict scoring data. Following are the prediction done using the tree model.

## 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

Let check whether person with ID 781,782,786,791 and 793 are eligible for loan or not. For person with ID=781, first will check his credit history is crirtical or not.If credit history is critical classify him as bad. if not we will check second condition. His credit history is all paid. So we will check second condition which is whether it is current or delay. If condition is statisfied check for employment if not classify him as good.here. As the credit history is not current or delay,he is classified as good.

For person with ID=782, he has credit history as current.so the first condition is not statisfied we will second condition that if credit history as current or delay this condition is true.Further we will check for if his employment is short or not.If employment is next condition is checked if not it is labelled as good. He has medium type of employment.Thus, we classify as good laon.

For person with ID=786, his Credit history is current so first condition is false and second is true.If second condition is true we will check for employment. If employment type is not “short” then it is classified as good.From the records, his employment status is unemployed so he will have credit standing as good.

For person with id=791, he/she has critical credit history hence he/she classified into bad loan.

For person with id=793,he has credit history as current as first condition is not statisfied will go to second condition whether it is current or delay as it is current, second condition is statisfied so next we will check his employment which is short, this condition is true so we proceed futher and again check whether credit history is current.from records he has current history so we can say the person with current credit history and short employement are categorize as bad loan.

To find the accuracy of the model,predictions are done on test data set. After prediction, confusion matrix table is shown below:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Bad Good  
## Bad 38 25  
## Good 12 83  
##   
## Accuracy : 0.7658   
## 95% CI : (0.692, 0.8294)  
## No Information Rate : 0.6835   
## P-Value [Acc > NIR] : 0.01452   
##   
## Kappa : 0.494   
##   
## Mcnemar's Test P-Value : 0.04852   
##   
## Sensitivity : 0.7600   
## Specificity : 0.7685   
## Pos Pred Value : 0.6032   
## Neg Pred Value : 0.8737   
## Prevalence : 0.3165   
## Detection Rate : 0.2405   
## Detection Prevalence : 0.3987   
## Balanced Accuracy : 0.7643   
##   
## 'Positive' Class : Bad   
##

Confusion matrix is table created between predicted values and actual values.We have actual values from our dataset and predicted values by the model using them we will decide how good the model performed. Here,references means actual values.We have two categories good and bad hence it is (2,2) matrix. There are 35 people who have a bad credit standing are correctly classified as bad this are also known as true positives. 87 people are actually good are identified as good credit staning also called as true negatives.Now, we will look at the errors,type 1 error is occurs when something is false while it is actually true. here, 8 have bad credit standing which are incorrectly identified as good.In second case, there are 28 who are good however predicted as bad this is called as type 2 error.

Sensitivity is measured true positive rate and specificity as true negative rate.The sensitivity of yhis model is better than its specificity.we can say that decision tree is good at classifing class label “good” than classifing label “bad”.Accuarcy of the decision tree is given as ratio of correctly predicted (either good or bad) to total number of the prediction.Accuarcy of this decision tree is 77.22%.

Now, we will improve the decision tree model by using two apporaches. First is ensemble techniques bagging and second is boosting method.

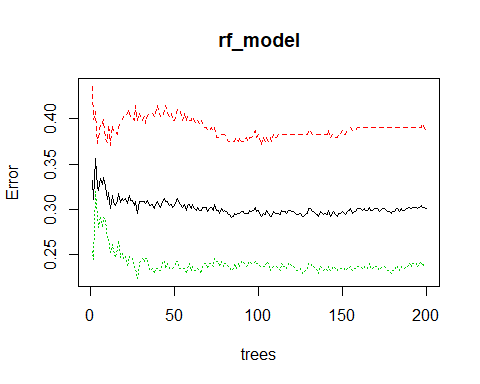
# Random Forest:

Bagging is technique in which the results of multiple decision trees (or any other models) are combined to get a generalized result. Here, we will be using random forest method for improvisation of decision tree.In this approach, a sampling technique is used to create subsets of observations with replacement from the original data.All subsets have same size as that of original set and selecetd randomly.Each subset is used for creating a decision tree therefore a large number of trees are generated. Every row or record is fed into these decision trees separately. For final result voting system is used for result.For example if majority of votes are for “good” label then final prediction result is “good”.

Random forest model:

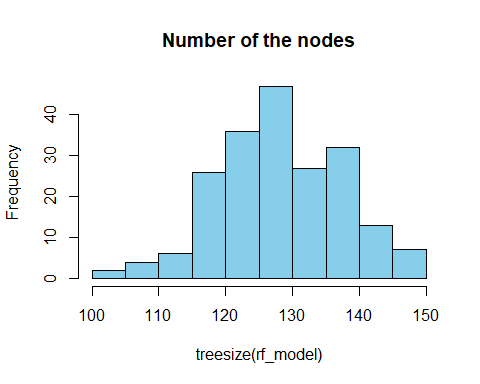
##   
## Call:  
## randomForest(formula = Credit.Standing ~ ., data = train, ntree = 200)   
## Type of random forest: classification  
## Number of trees: 200  
## No. of variables tried at each split: 3  
##   
## OOB estimate of error rate: 30.06%  
## Confusion matrix:  
## Bad Good class.error  
## Bad 157 99 0.3867188  
## Good 88 278 0.2404372

Here, total of 500 trees are build by default. But we print that model after 200 trees the error estimation remains constant hence we will use number of trees as 200. Number of variable tried at each split is alway taken as square root of number of variables hence as there are 13 predictors it is taken as 3.We can change this but as it is giving better accuracy at 3 so it is not changed.Next, is out-of-bag error rate,it is error estimate for the cases which were not used while creating the tree. Less the OOB error,better the accuracy of random forest model.



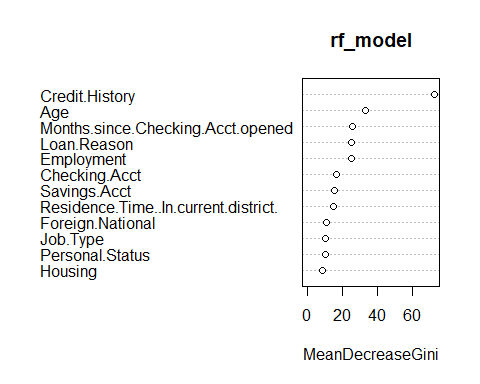
Here, we can see that the error is decreased as the number of trees increases.But after 150 the error rate is mostly constant. Hence we are using 200 number of trees in this model.

Number of nodes in tree:



The above graph shows histogram of tree size in this model. It is clear from graph that a large number of trees have nodes between 120 to 140.

Important variables:



Credit history is most important variable used for prediction. Followed by Age. Months since account opened ,loan reson and checking history have nearly same importance.

Predictions and accuracy of the model:

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Bad Good  
## Bad 46 15  
## Good 17 80  
##   
## Accuracy : 0.7975   
## 95% CI : (0.7263, 0.8572)  
## No Information Rate : 0.6013   
## P-Value [Acc > NIR] : 1.145e-07   
##   
## Kappa : 0.5753   
##   
## Mcnemar's Test P-Value : 0.8597   
##   
## Sensitivity : 0.7302   
## Specificity : 0.8421   
## Pos Pred Value : 0.7541   
## Neg Pred Value : 0.8247   
## Prevalence : 0.3987   
## Detection Rate : 0.2911   
## Detection Prevalence : 0.3861   
## Balanced Accuracy : 0.7861   
##   
## 'Positive' Class : Bad   
##

Test data is fed in for testing the performance of the model.There are 80 true positives i,e credit standing predicted as good which are actually good.Important thing in this model is that specificity which is true negative rate is higher compared to the decision tree model.It is better in classifing bad labels than the decision tree model.Hence, type 1 error number is also reduced.Accuracy of the model which is increased than previous model.

The model is giving better results because in the random forest method a large number of tree are created using different subsets of the data(in this case 200 trees were created).So, even if one tree does wrong prediction another tree might predict it correctly. Output from each tree is taken into consideration for example if majority of trees predict as good loan then final output is good loan.This improves the accuracy of the result.

# Adaboost method:

Second method used for improving the deisicion tree is boosting.In boosting a number of weak learner are combined to form a strong model.This is sequential process. We will use adaboost method of boosting and let see how it works.

In this model, first the data is divided into training subset from the original dataset.All points are given same weight initially.A base model is created and used for predicting that dataset.then error is calculated using predicted and actual values.The weight of observations which are incorrectly predicted is increased.Considering error of previous model new model is created and this process continues generating multiple models, each correcting the errors of the previous model.The final result is weighted output of all the models(base learners).

## adaboost(formula = Credit.Standing ~ ., data = train, nIter = 70)  
## Credit.Standing ~ .  
## Dependent Variable: Credit.Standing  
## No of trees:70  
## The weights of the trees are:0.80279220.69342960.66704820.58840360.51124050.49365010.45956180.4129380.40186720.41972880.37159680.41708260.30931240.3527250.39405680.33953610.33256990.31837720.24568490.23299310.2985920.22289630.20958130.23554860.21390340.18557710.21660230.23998620.241680.20628670.21168680.21642250.16821110.19109280.15252820.18994960.16747470.21459370.15146050.1705140.14835870.16194830.10701120.096963970.11248930.10890760.110760.10274740.10394340.11187750.10898250.1155050.10911950.10014420.088189880.10524980.10787190.10446640.10911710.10021180.11084560.092952570.083141920.069213940.069917450.10228030.09514070.068084250.077351960.09182988

Here, number of iteration used is 50 by default. It can be adjusted accordingly.70 iternation are sufficient for this model to give better result. Last is the weights of all the trees created is displayed.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Bad Good  
## Bad 39 15  
## Good 24 80  
##   
## Accuracy : 0.7532   
## 95% CI : (0.6784, 0.8182)  
## No Information Rate : 0.6013   
## P-Value [Acc > NIR] : 4.25e-05   
##   
## Kappa : 0.4725   
##   
## Mcnemar's Test P-Value : 0.2002   
##   
## Sensitivity : 0.6190   
## Specificity : 0.8421   
## Pos Pred Value : 0.7222   
## Neg Pred Value : 0.7692   
## Prevalence : 0.3987   
## Detection Rate : 0.2468   
## Detection Prevalence : 0.3418   
## Balanced Accuracy : 0.7306   
##   
## 'Positive' Class : Bad   
##

Similar to the random forest model, the true negative rate(specificity) is increased in this model. Type 1 error is increased but type 2 error is decreased as compared to the decision tree. Accuracy of the model is improved compared to decision tree model.

This model gives better results because this model takes in to account the mistakes made by the previous tree and focus on classifing that obervations correctly.Further, perforamance of each tree also called as stumps is taken into consideration while final voting. The result of the adaboost combination of the weighted output of all the weak learner or stumps.

# ANAMOLY DECTECTION:

Anomaly detection is the identification of observations that are suspicious as they differ significantly from the majority of the data. The company has used a process that is a mixture of a grading system and human input to grade each past loan as good or bad.So we will check is there any particular time period in which system performed poorly. For that we will first predict the whole dataset using our tree model.This predictions along with the actual result of the credit standing that we have in the dataset, we will put in a data frame. The ID’s from original dataset are added in this data frame. Next, we will calculate the error in the prediction. Error value is assigned as 1 if the observation is incorrectly predicted otherwise 0. Last is to check if there are any consecutive 1’s in error column (using loops).

Observations from ID number 623 to 637 have all predicted values as good loan and actual value bad loan.This the time period where system might have some error because of that it has given value bad for the people who should have credit standing as good.

# Conclusion:

In conclusion, from all three models credit history plays vital role in classifing loans as good or bad. Ensemble learning techniques can be used for building better models with high accuracy.