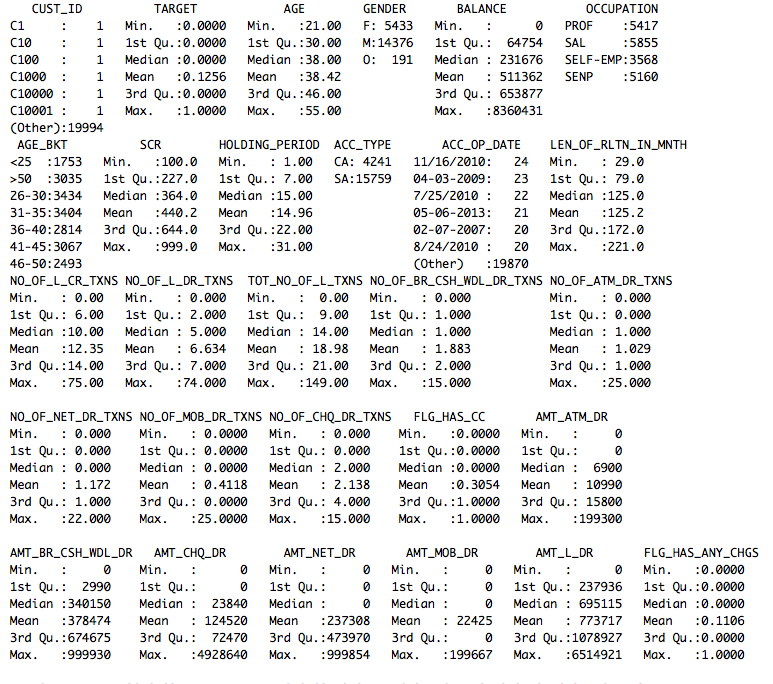
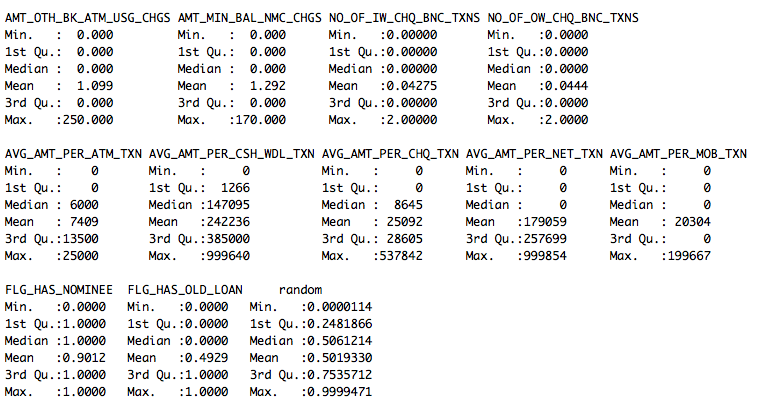
Import the data file

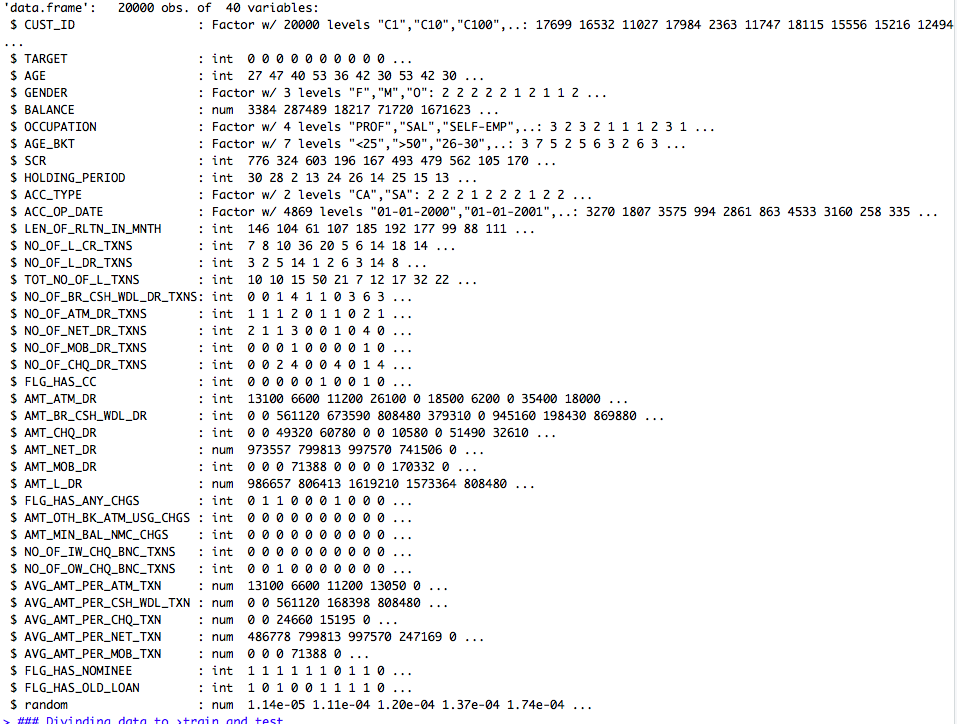
Screen Shot 2018-09-08 at 9.13.12 AM.png

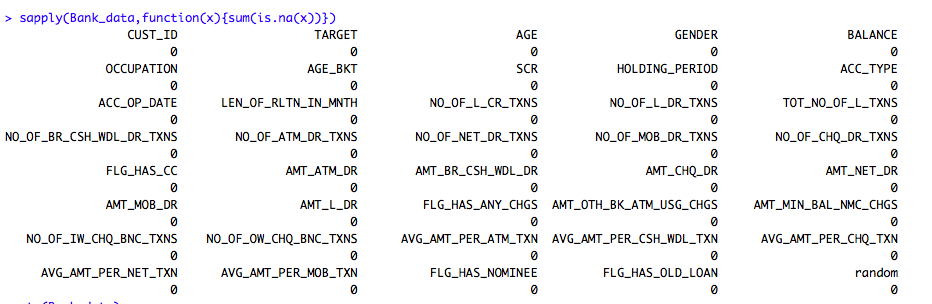
Perform univariate analysis and understand data

Screen Shot 2018-09-08 at 9.13.42 AM.png







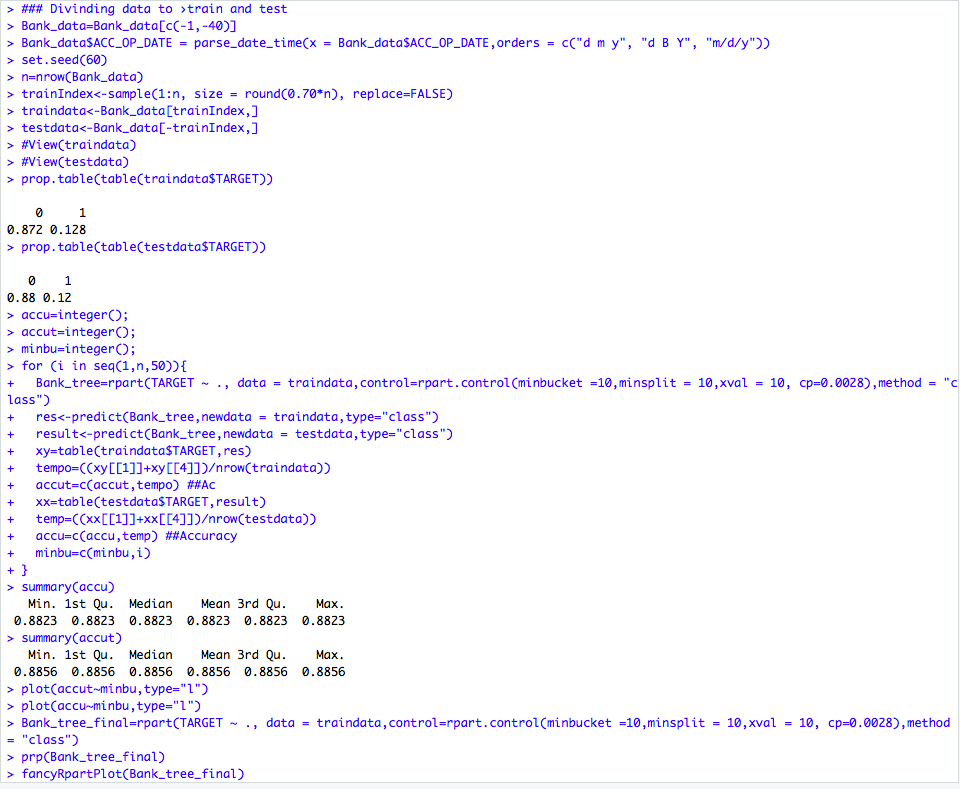


We can observe that there are no rows which have any missing values. And after checking all the data, we have noticed that date is not in date format, we have a column with sequence which must be eliminated and random value column which must be deleted.

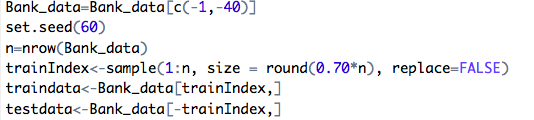
We can perform decision tree model below scenario

1. Removing Random and ID column and change date to date data type.
2. Reduce dimensions with co-linearity, scale data and build model.

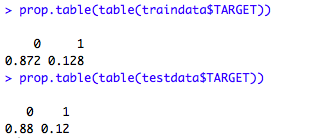
Removing Random and ID column and change date to date data type.



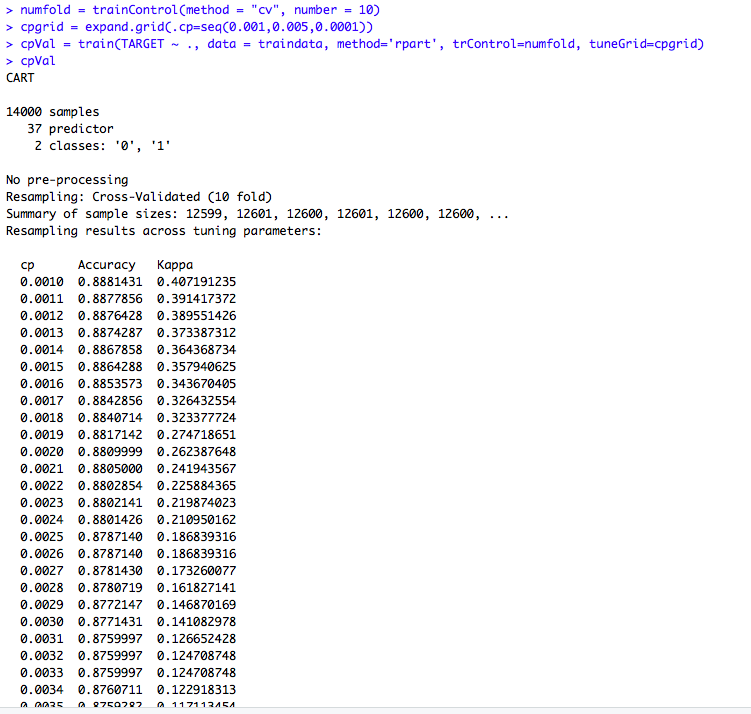
Divide data to train and test data.

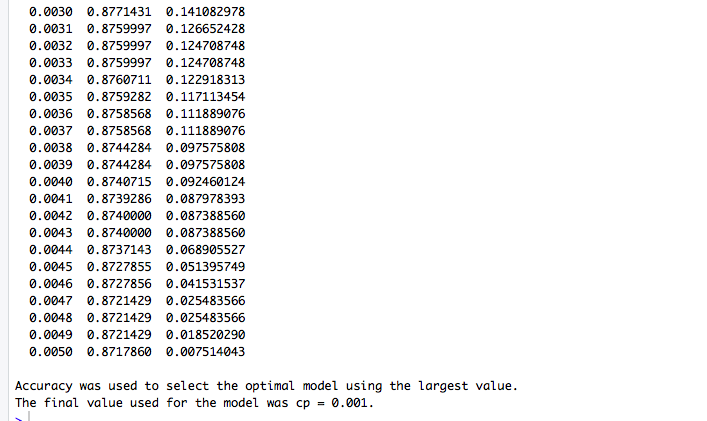


Check the data proportion.

Screen Shot 2018-09-08 at 9.14.09 AM.png

Choosing appropriate CP value.



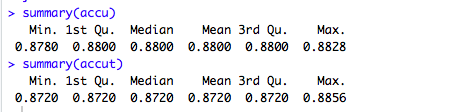


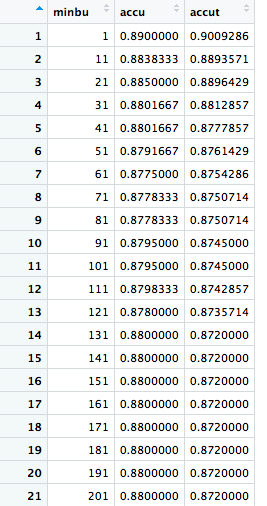
0.001 is the purest form and the decision tree obtained would be over fit. The decision tree obtained using 0.19 is very complex hence 0.20 to 0.28 can be selected.

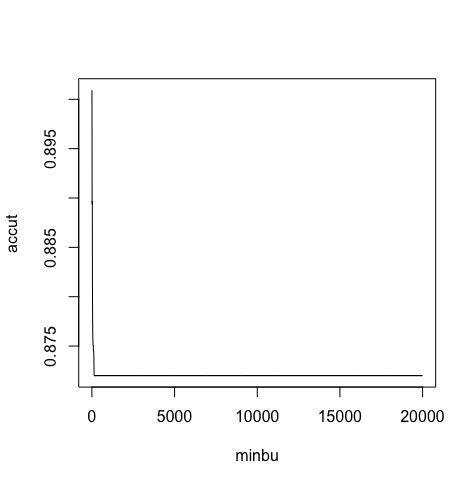
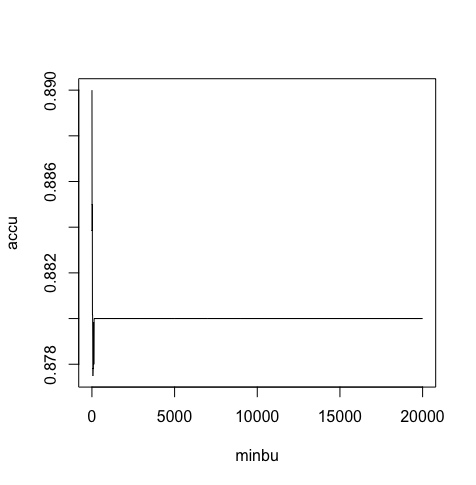
Build decision tree in loop to find the best minimum bucket.



Plot min bucket graph.

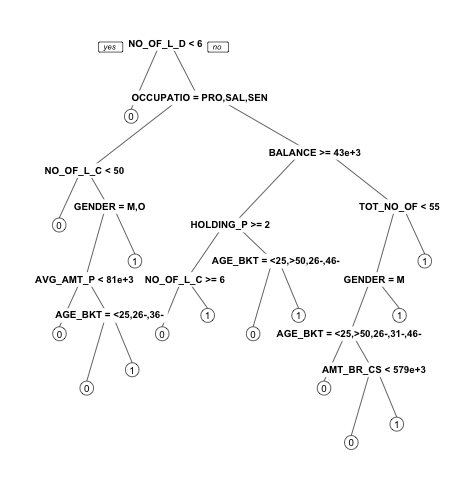






20 can be chosen as min bucket.

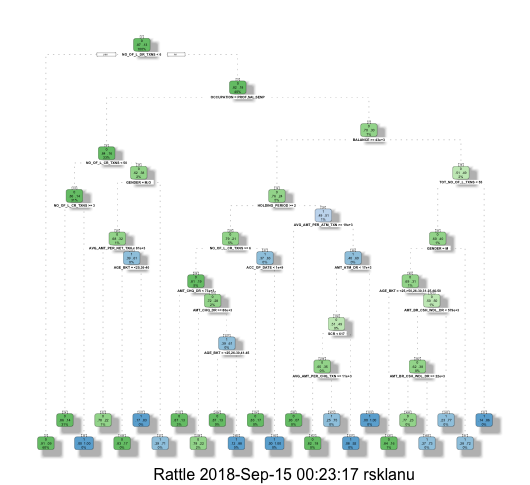
Build final model with min bucket with 0.28 CP and 20 as min bucket.

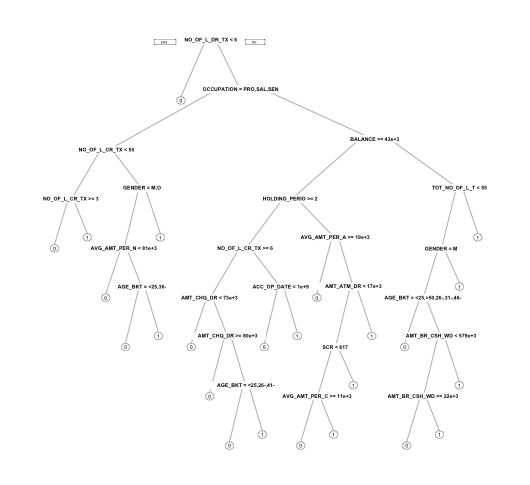


Build tree with min bucket 20 and CP 0.0028.

Screen Shot 2018-09-15 at 1.35.38 AM.png

Visualize the tree





We were able to visualise the above decision tree with the model built.

For the above model even through the accuracy was 88-89%

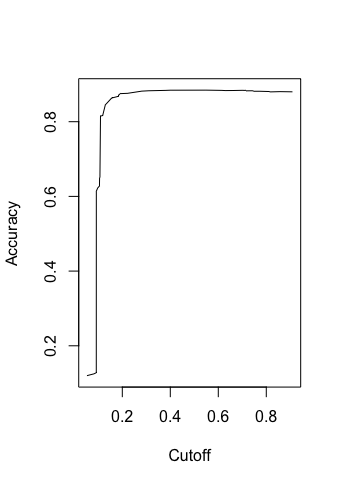
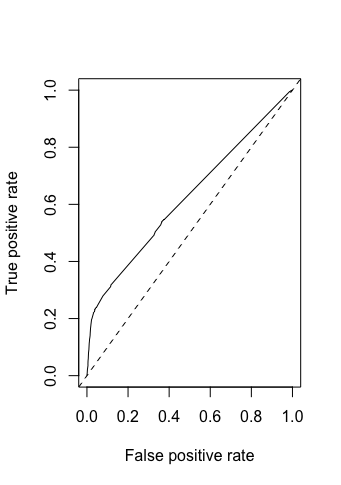
Confusion matrix looks like below.



Confusion matrix for min bucket 10 and cp=0.2 is given below.

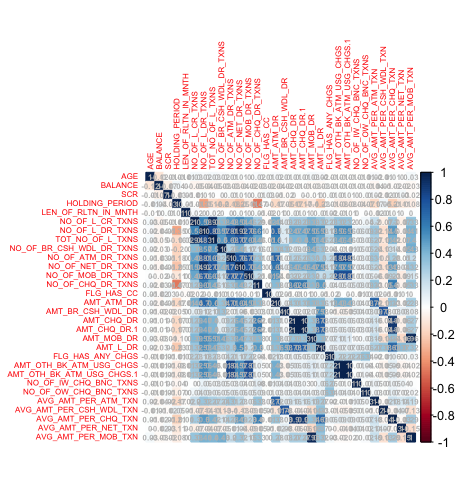


ROC and Accuracy plot looks like below

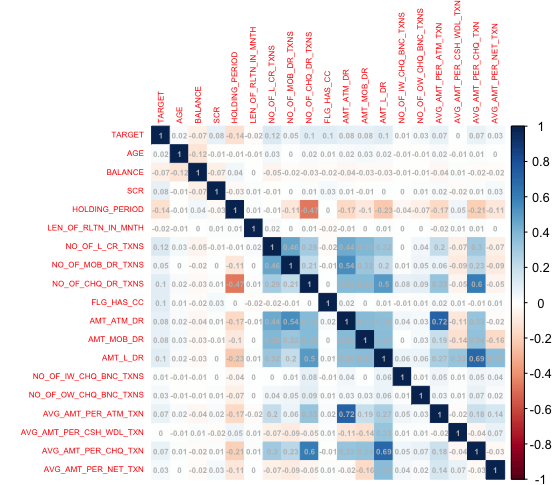


We then tried to reduce collinear dimensions and scaling data, we could arrive at below model

Finding collinearity.



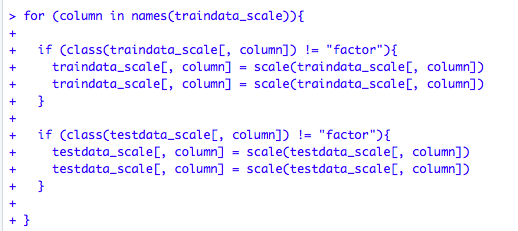
Reduced columns coliniaerity



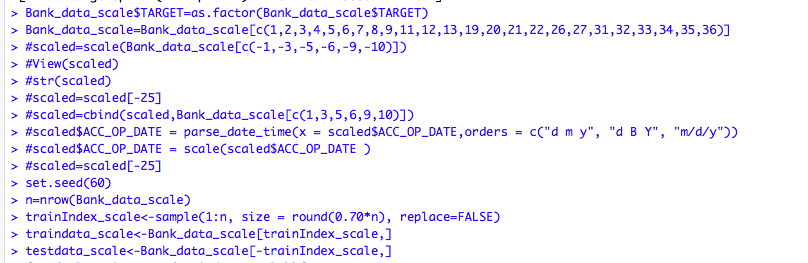
Reading data from excel and removing ID and random column



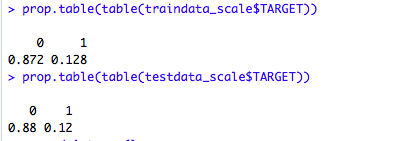
Scaling data after removing all colinear row which provide similar data. Also date has been delete as it gives same info as LEN\_OF\_RLTN\_IN\_MNTH.

Dividing data to train and test

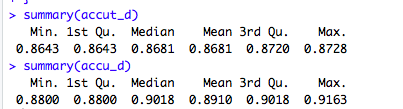
Dividing data for train and test and removing unwanted columns.



Checking proportion and finding min bucket



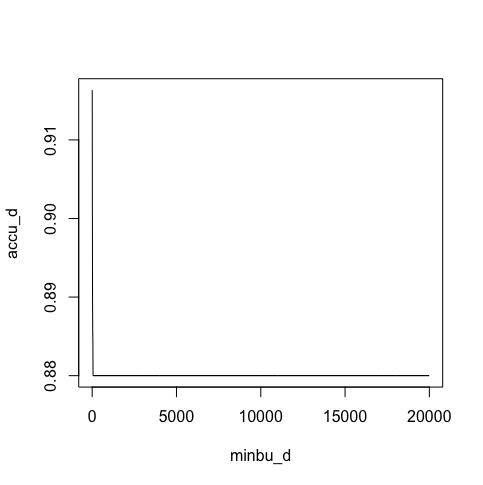
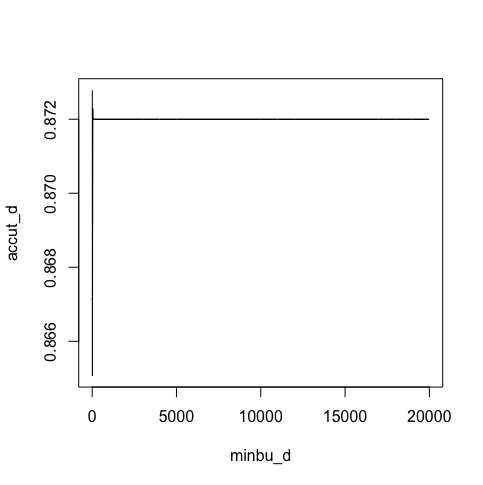
Plot accuracy and find min bucket with max accuracy in both test and train.



When we compare the accuracy and min bucket bets case is overfit model whose bucket size is one. To avoid over fitting and get better accuracy I have chosen bucket size of 20.

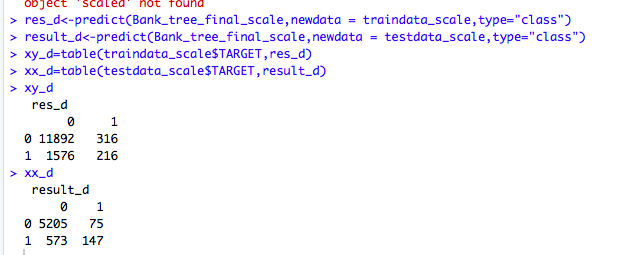
Here is where we can arrive at median accuracy of both test and train data.

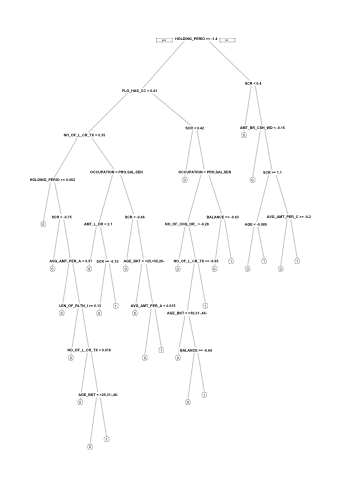
On the training set we have got approximately 87% but for test we have got 91% accuracy.

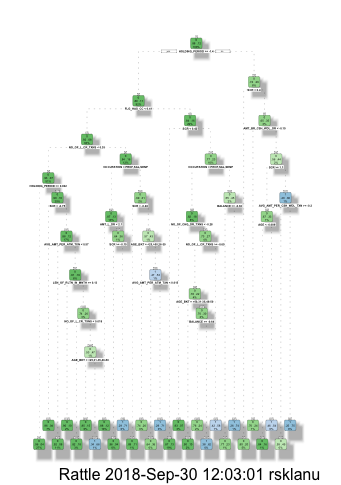


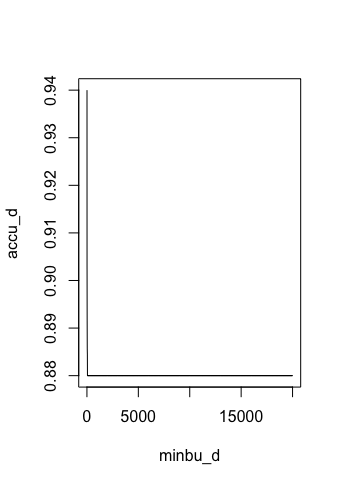
Building final model

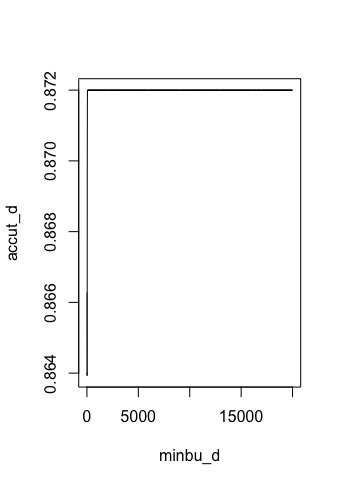




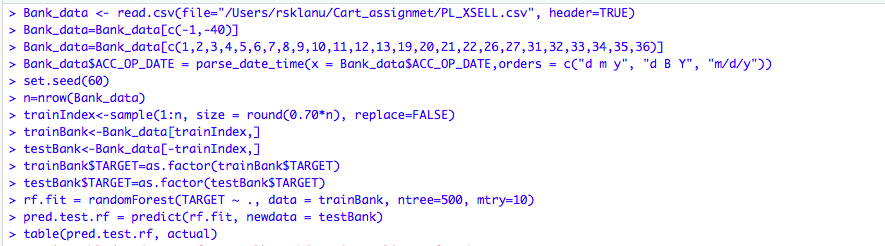
Visualising the tree

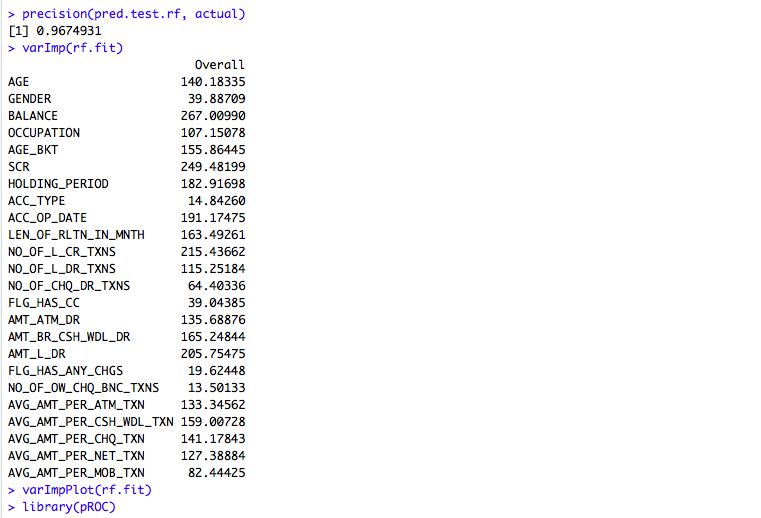
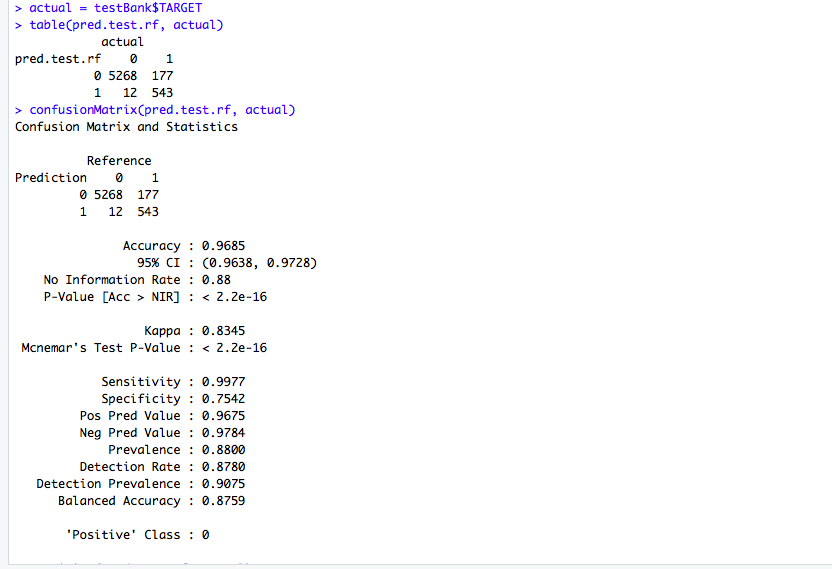


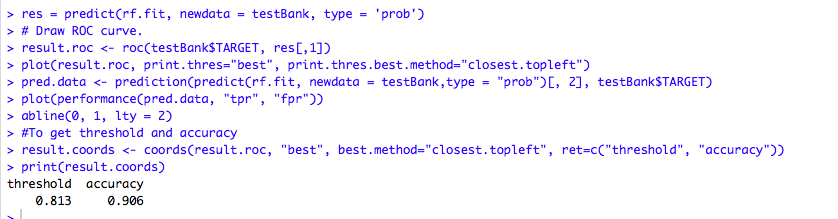


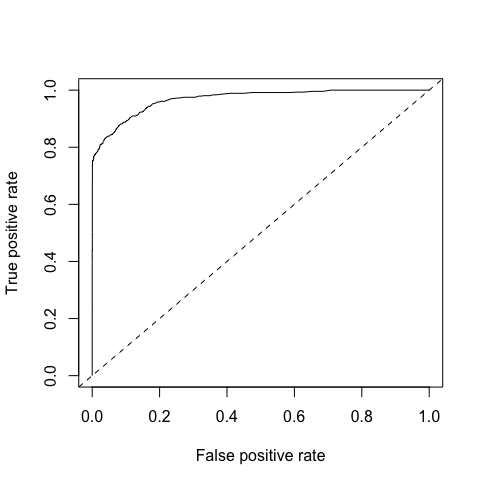
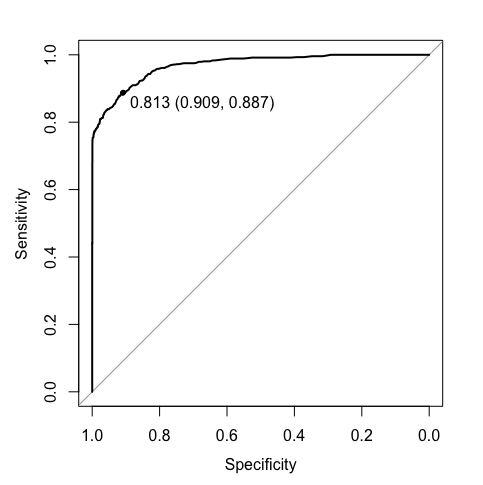
ROC and AUC curve

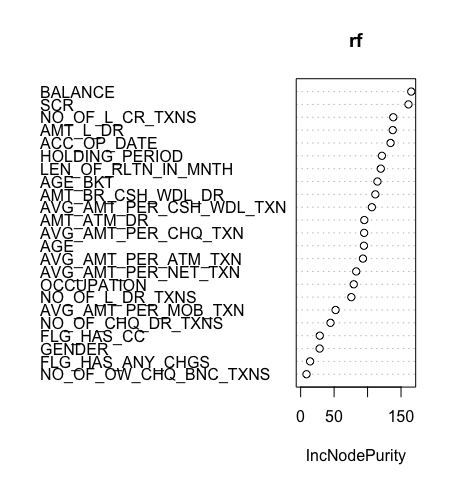
Using Random forest.



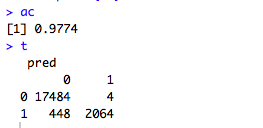






Significant variables in random forest are plotted as below.

Accuracy achieved using random forest model is 98%



Ratio of misclassification is less compared to decision tree model.

The only disadvantage of using random forest is , its a black box approach, we are unable to explain why we classified a customer to defaulter or non defaulter.

I therefore considered top 20 , non related columns from the above graph and tried to draw a decision tree and check if better results can be obtained using decision tree, I obtained the similar result.

Based on the requirement if the customer requires an explanation or not we ca use either of the methods can be used.