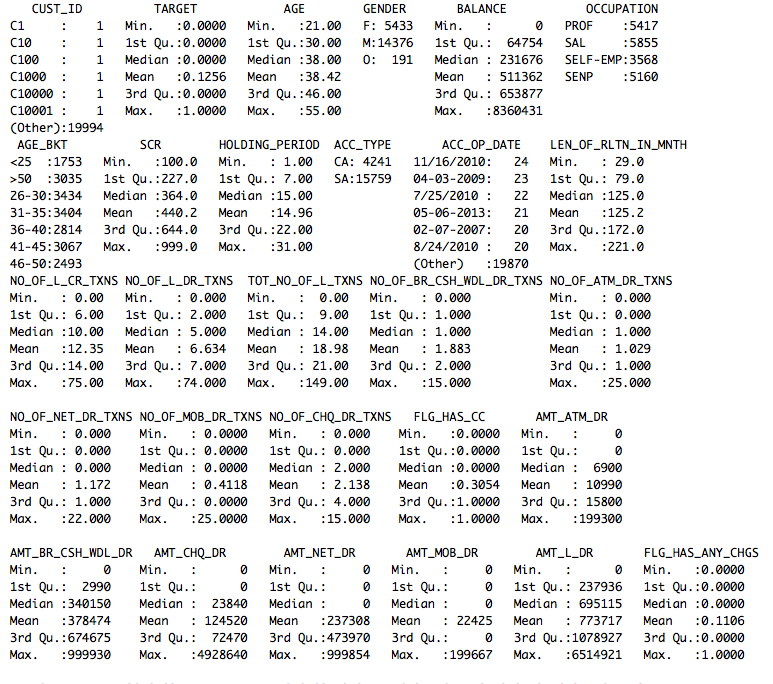
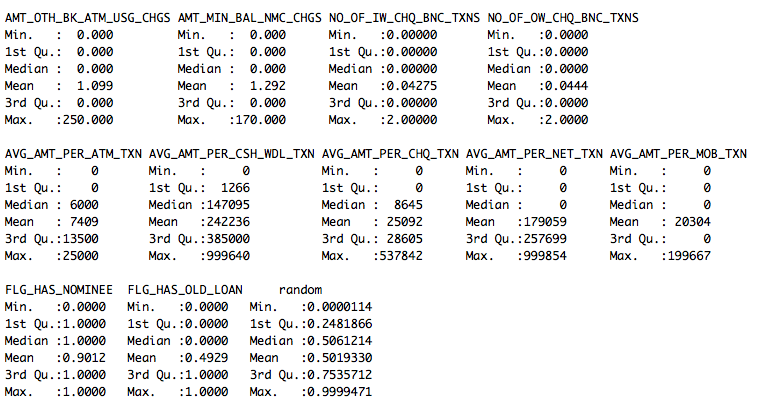
Import the data file

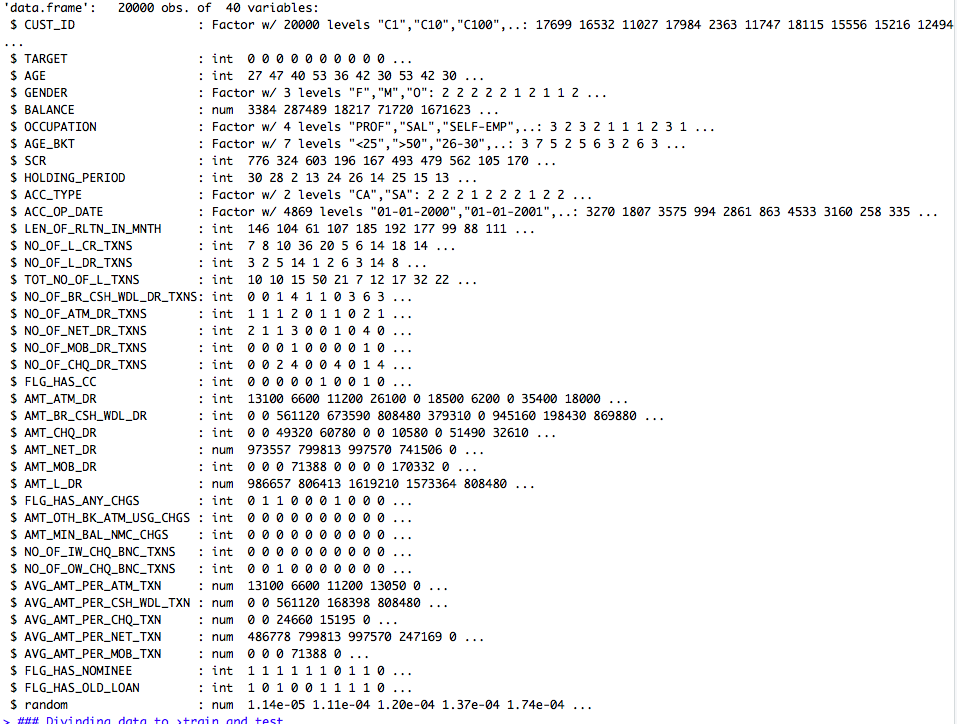


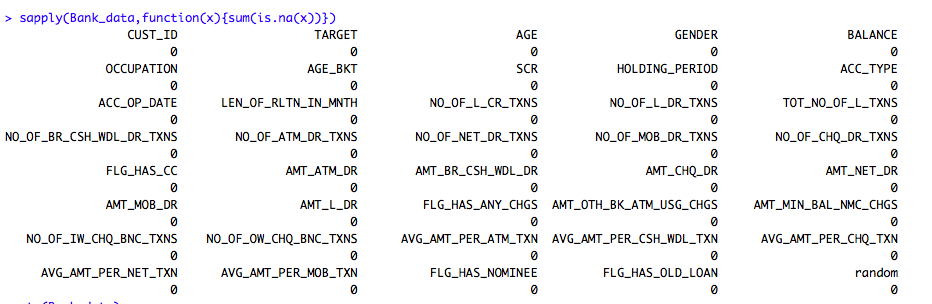
Perform univariate analysis and understand data









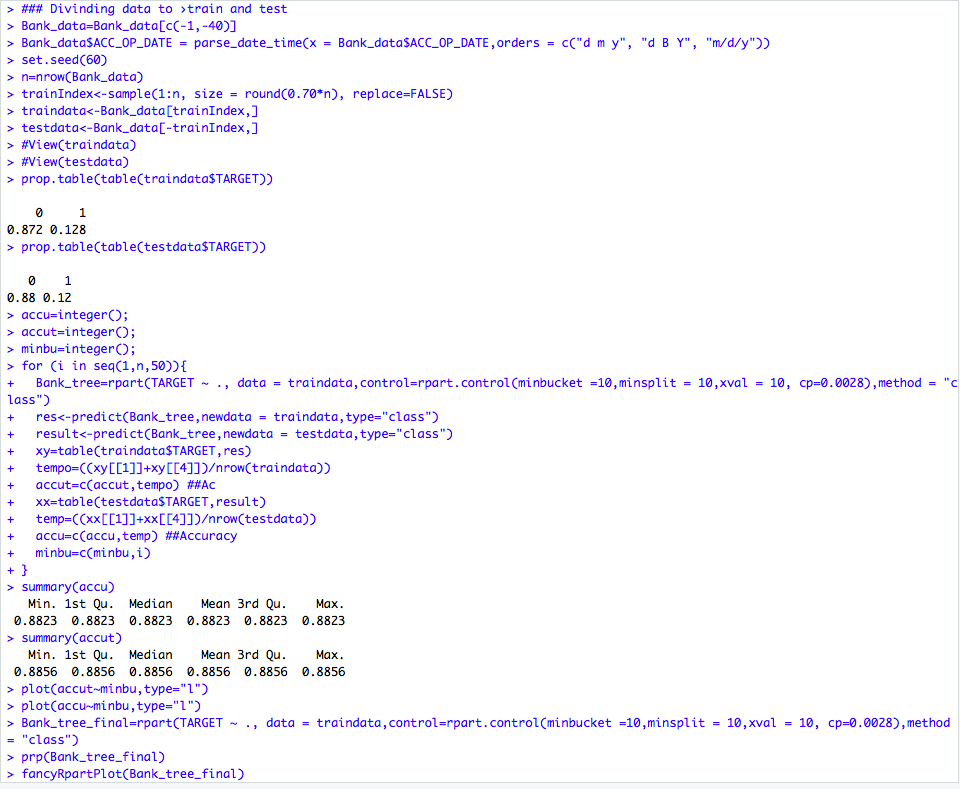


We can observe that there are no rows which have any missing values. And after checking all the data, we have noticed that date is not in date format, we have a column with sequence which must be eliminated and random value column which must be deleted.

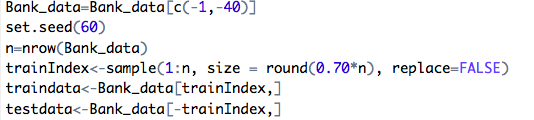
We can perform decision tree model below scenario

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

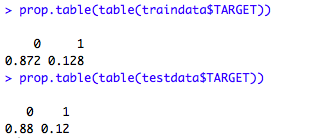
Removing Random and ID column and change date to date data type.



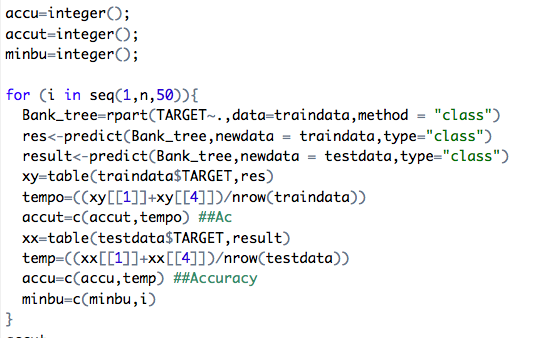
Divide data to train and test data.



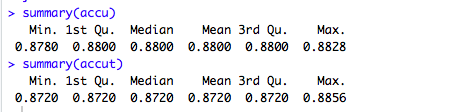
Check the data proportion.

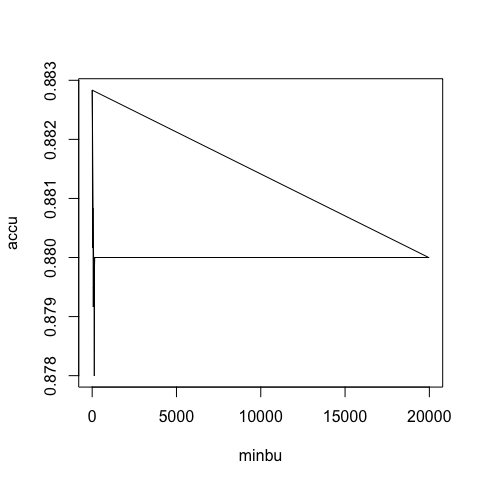
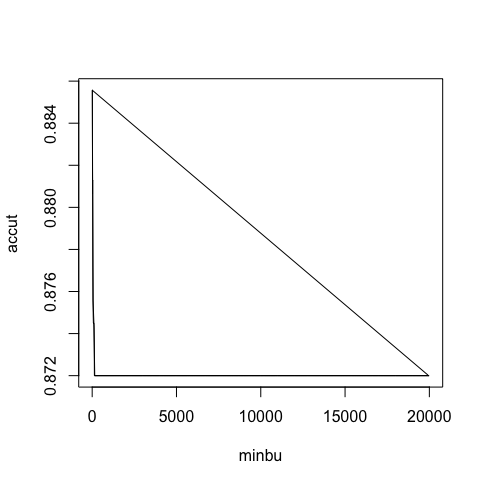


Build decision tree in loop to find the best minimum bucket.



Plot min bucket graph.

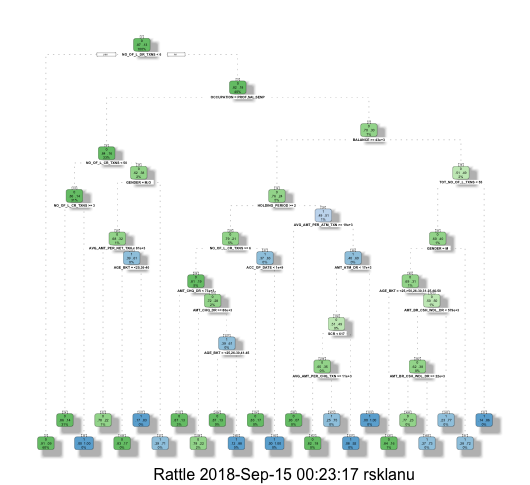


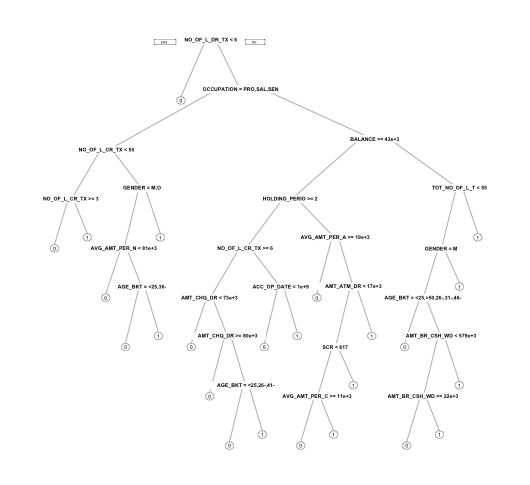


Build final model with min bucket with maximum accuracy, in this case it doesn’t matter as all buckets have same accuracy.



Visualize the tree

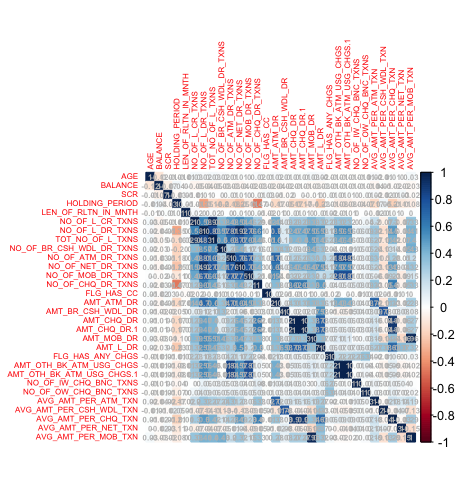




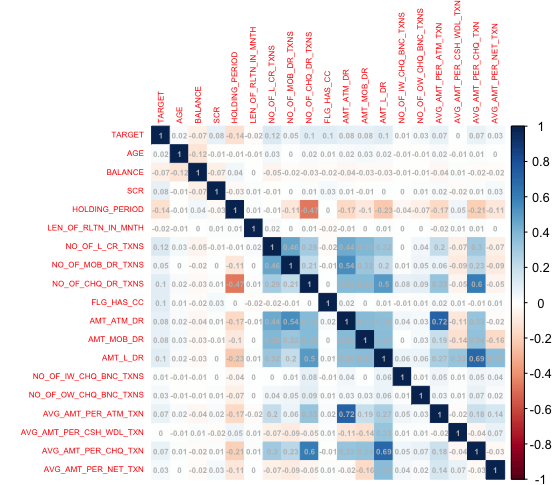
We were able to visualize the above decision tree with the model built.

After reducing colinear dimensions and scaling data, we could arrive at below model

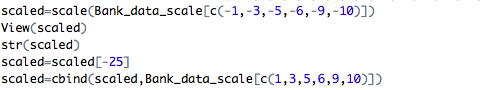
Finding collinearity.



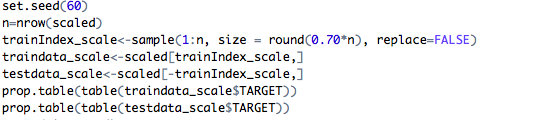
Reduced columns coliniaerity



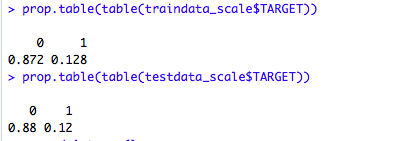
Scaling data after removing all colinear row which provide similar data. Also date has been delete as it gives same info as LEN\_OF\_RLTN\_IN\_MNTH.



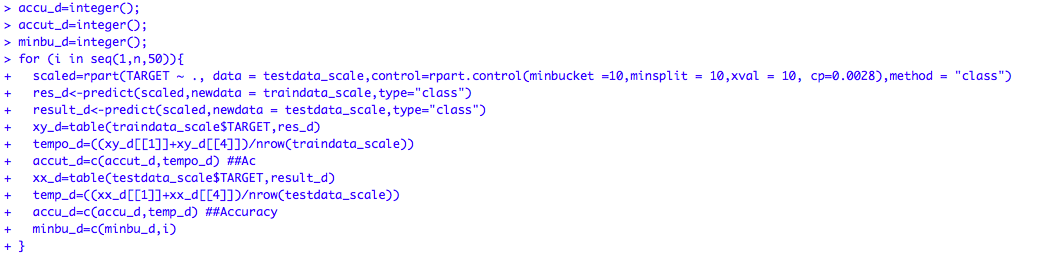
Dividing data to train and test

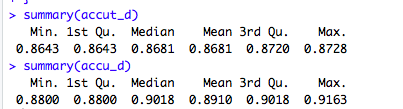


Checking proportion and finding min bucket



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

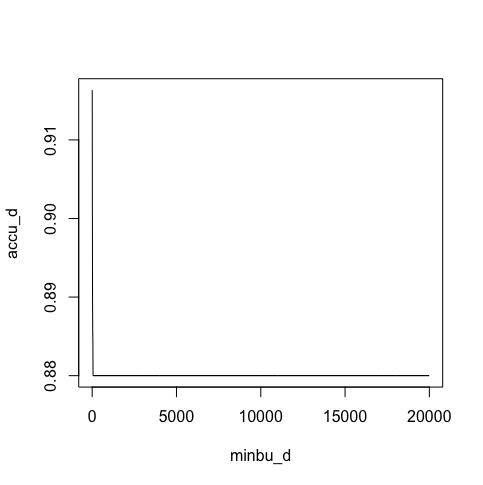
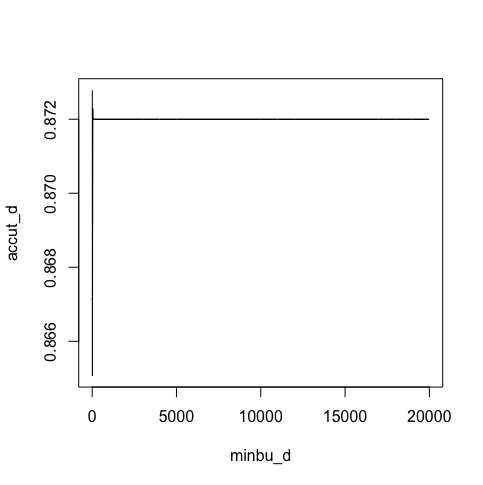




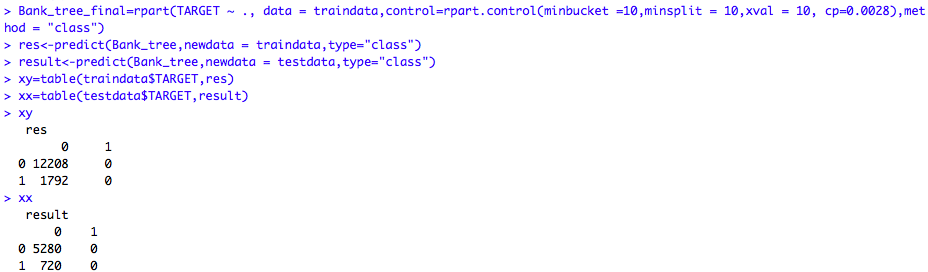
When we compare the accuracy and min bucket bets case is overfit model whose bucket size is one. To avoid over fitting and get better accuracy I have chosen bucket size of 10.

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.

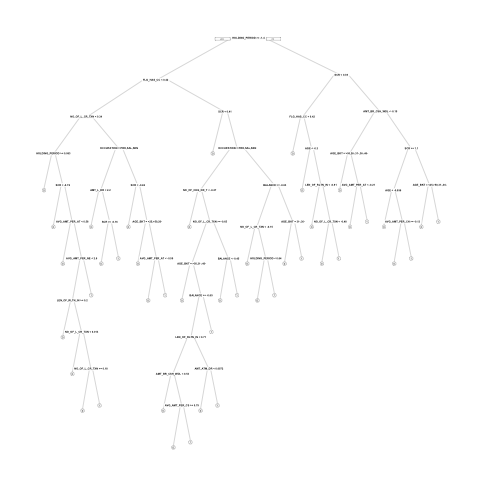


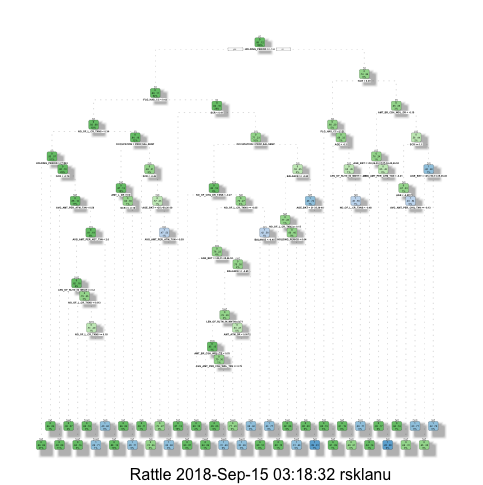
Building final model

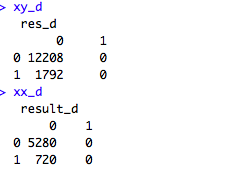


Sensitivity is 0, No value is classified target as 1 hence this model cannot be considered

Plotting final decision tree







As the sensitivity for the model is important and we see this model is not classifying any value to category one this may not be the ideal model, even though it has high accuracy rate.