GERMAN CREDIT CARD FRAUD DETECTION
DECISION TREE ANALYSIS

IDS 572
Assignment 1
February 2018

Kalindi Deshmukh

Q1:

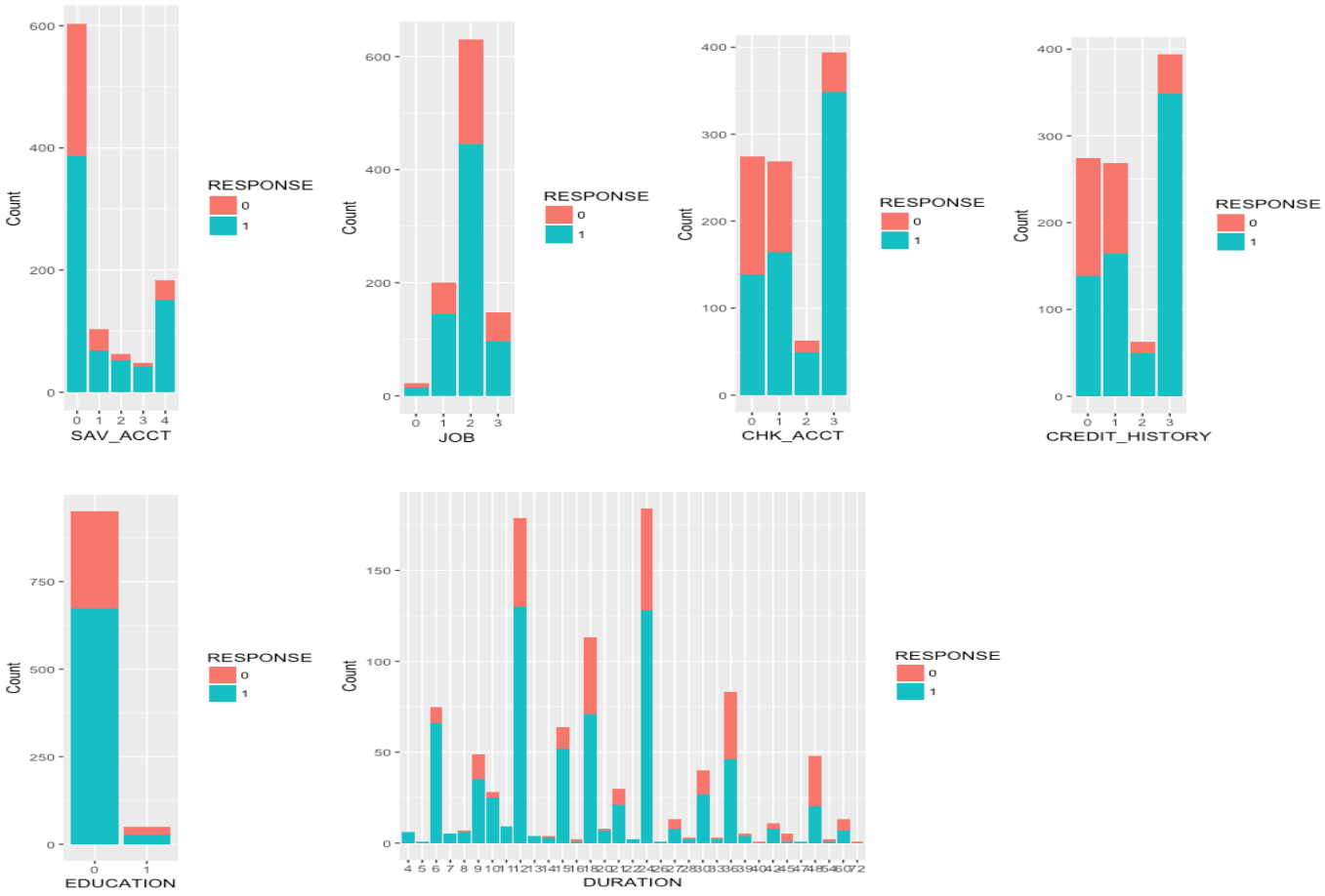
Explore the data:

The German Credit dataset has data on 1000 past credit applications. The number of “Good” cases that encoded as 1 in Response variable is 700 and the number of “Bad” cases that encoded as 0 in Response variable is 300. The proportion of “Good” to “Bad” cases are 700/300 that the result is 7:3.

There are some missing values in column W that is Age. We can use method “mean” to fill in missing values in this column. Also, for some columns there are NA values that we remove them and use zero instead of NA.

CHK_ACCT, JOB, SAV_ACCT, PRESENT_RESIDENT, CREDIT_HISTORY, DURATION, and EDUCATION are very important for the outcome of interest. Because based on our dataset most people with good conditions of these variable have good credits.

Based on some below plots, some of bad credit cases are more prevalent in certain value ranges of specific variables. For instance, most people who don’t have any saving account are as bad credit or people who has no education are bad clients. This is what we expected.



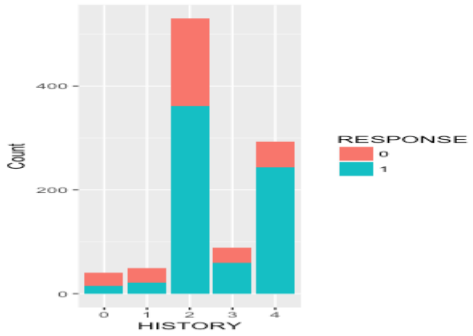
In 300 bad clients, 278 of them (92%) have no education. Also, 217 of them (72%) have less than 100DM in average balance in their savings account. Moreover, 135 of them(45%) have checking account and in categorical type (0: <0 DM).

The scatter plot between age and amount shows people less than aged 40 and their credit amount is less than 10000 are part of bad clients.

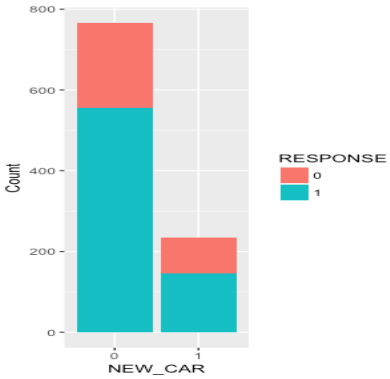
The scatter plots between age and duration shows people less than age 40 and their duration of their credit in months are more than 35 are part of bad clients.

Plot of some important variables:

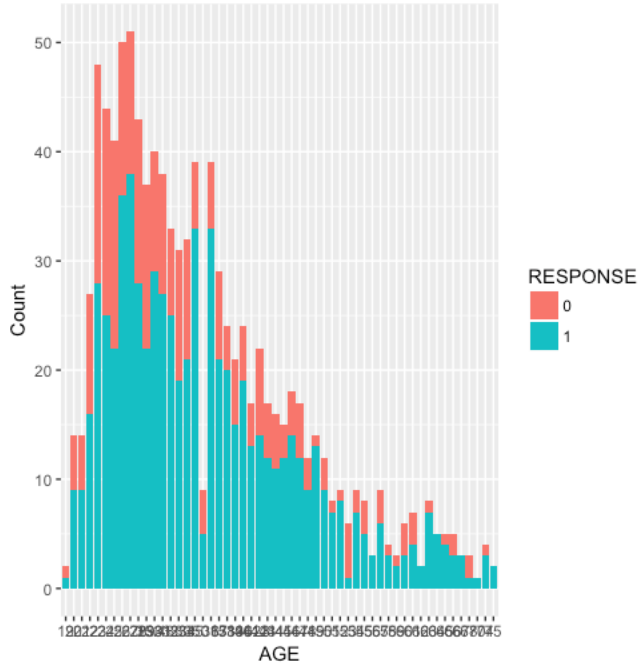
- Based on HISTORT plot Most of clients are part of “existing credits paid back duly till now”, also most of people with bad credits are part of same group



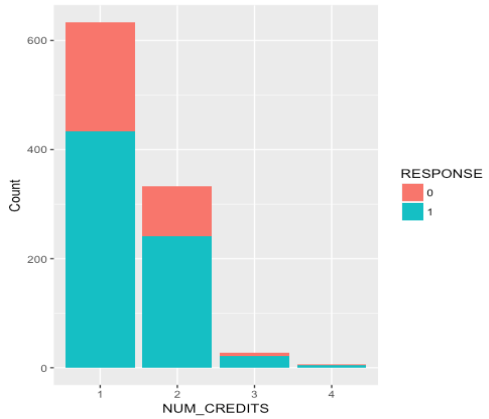
- Based on NEW_CAR plot a quarter of clients have a new car and around 10% ARE BAD CLIENTS



- Based on AGE plot most of bad clients are between 21- 37



- Based on NUM_CREDIT plot most of bad clients have one or two existing credits at this bank



Numerical variables

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
DURATION	4.0	12.0	18.0	20.9	24.0	72.0
AMOUNT	250	1366	2320	3271	3972	18424
AGE	19.00	27.00	33.00	35.48	42.00	75.00
NUM_CREDITS	1.000	1.000	1.000	1.407	2.000	4.000
NUM_DEPENDENTS	1.000	1.000	1.000	1.155	1.1000	2.000

Categorical Variables:

variable	0	1	2	3	4	mode
CHK_ACCT	274	269	63	395		3
HISTORY	40	49	530	88	294	2
NEW_CAR	766	234				0
USED_CAR	897	103				0
FURNITURE	819	181				0
RADIO.TV	720	280				0
EDUCATION	950	50				0
RETRAINING	903	97				0
SAV_ACCT	603	103	63	48	183	0
EMPLOYMENT	62	172	339	174	253	2
INSTALL_RATE	0	0	0	973	0	3
MALE_DIV	950	50				0

MALE_SINGLE	**452**	**548**				**1**
MALE_MAR_or_WID	**908**	**92**				**0**
CO.APPLICANT	**959**	**41**				**0**
GUARANTOR	**948**	**52**				**0**
PRESENT_RESIDENT	**130**	**308**	**149**	**413**		**1**
REAL_ESTATE	**718**	**282**				**0**
PROP_UNKN_NONE	**846**	**154**				**0**
OTHER_INSTALL	**814**	**186**				**0**
RENT	**821**	**179**				**0**
OWN_RES	**287**	**713**				**1**
JOB	**22**	**200**	**630**	**148**		**2**
TELEPHONE	**596**	**404**				**0**
FOREIGN	**963**	**37**				**0**
RESPONSE	**300**	**700**				**1**

Q2:

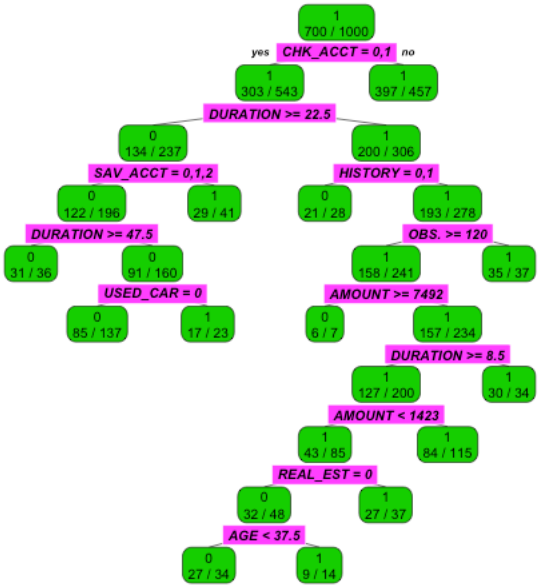
Decision tree is developed on the full data. Also, **Gini Index** and **Information** is considered . **Gini Index** provides the most accuracy on our data.

Variable importance:
CHK_ACCT - DURATION - HISTORY - AMOUNT SAV_ACCT REAL_ESTATE - USED_CAR - RADIO.TV - AGE - JOB - PRESENT_RESIDENT- PROP_UNKN_NONE - GUARANTOR MALE_MAR_or_WID - EMPLOYMENT - INSTALL_RATE

Gini Index			information	
	0	**1**	**0**	**1**
0	170	72	137	64
1	130	628	163	636

Accuracy	
Gini Index	0.798
information	0.773

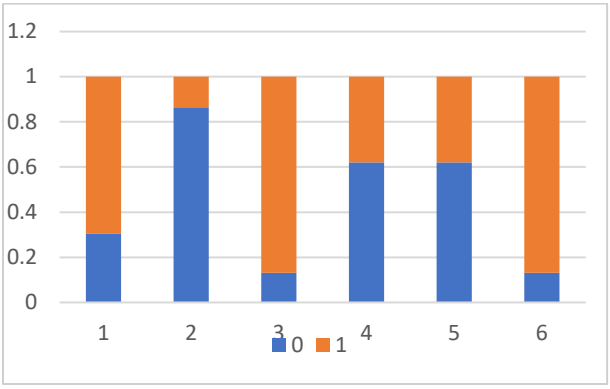
Decision tree split= GINI INDEX



Different performance measures were considered such as recall, sensitivity. Also, We considered lift, ROC , and AUC. A model is reliable when there is few changes. We can check reliability of our models after making training and testing sets. However, the model is almost robust.

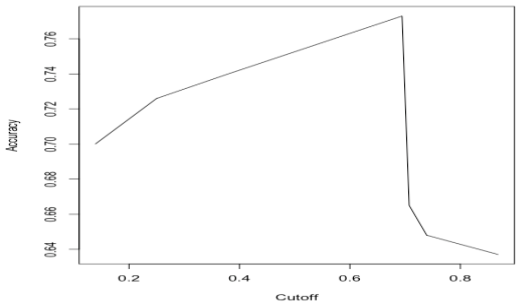
Lift Chart:

	0	1
1	0.3057554	0.6942446
2	0.8611111	0.1388889
3	0.1312910	0.8687090
4	0.6204380	0.3795620

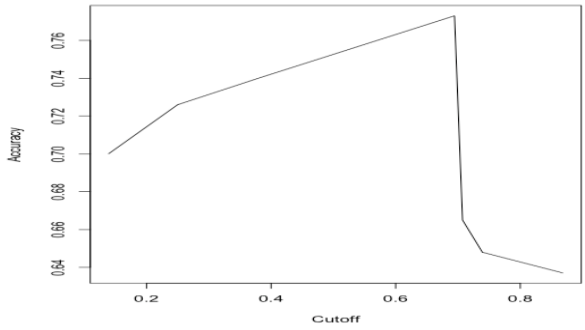


5	0.6204380	0.3795620
6	0.1312910	0.8687090

Lift:

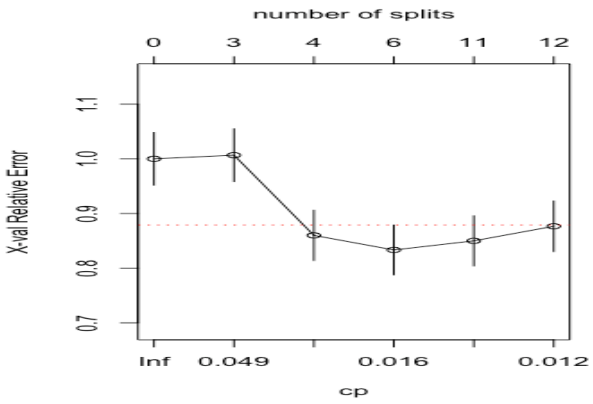


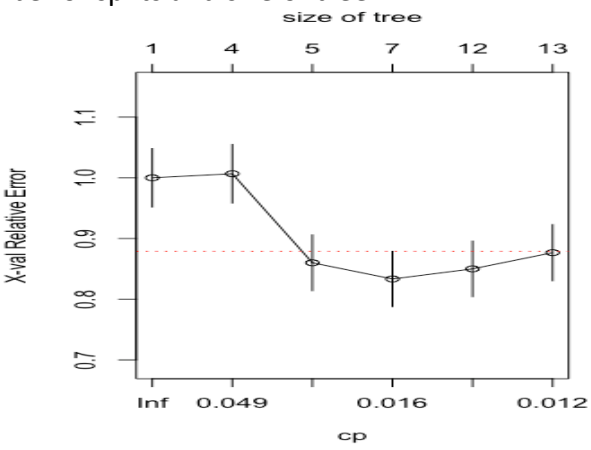
Accuracy:



plots of complexity parameter:

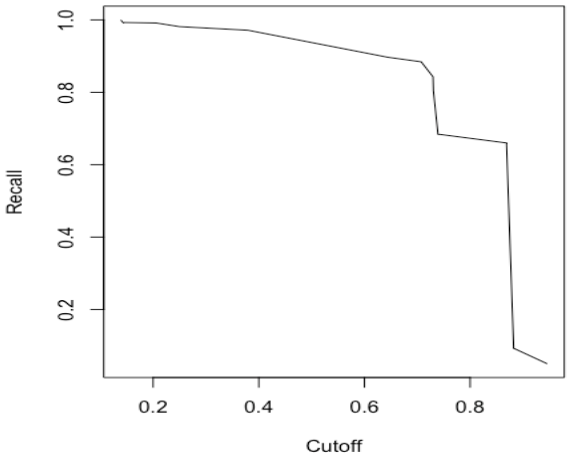
Below plots show that we consider our model with different number of splits and size of tree. Also, we can see in plot the minimum amount of errors for specific number of splits and size of tree.

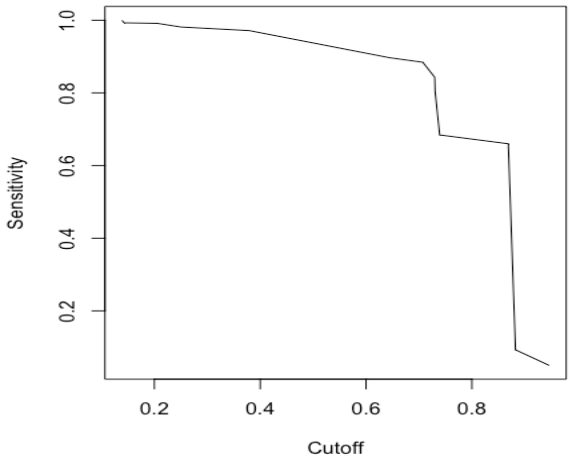




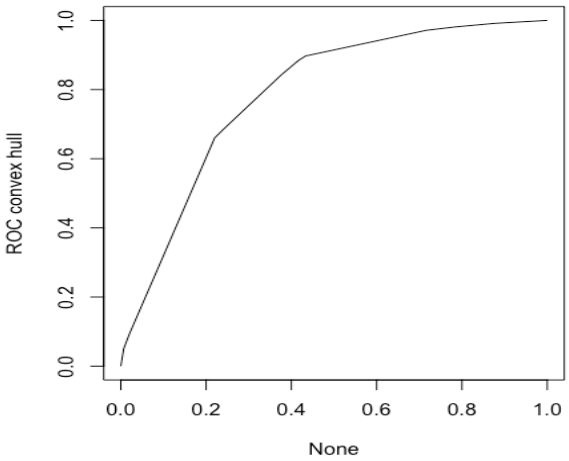
Performance: recall

performance: sensitivity





ROC:



A binary decision tree:

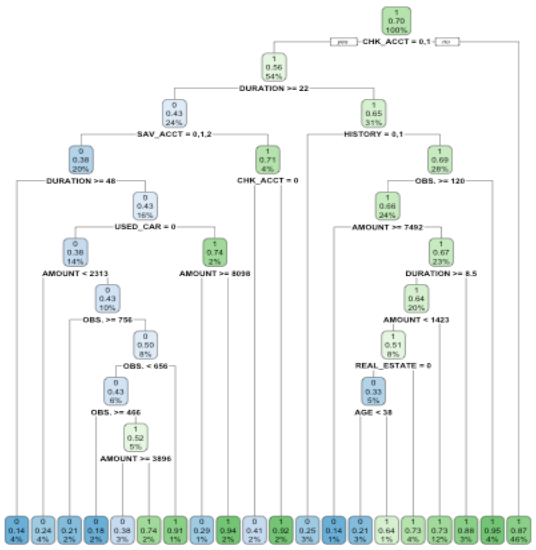


Table of complexity parameter:

	CP	nsplit	rel error	xerror	xstd
1	0.05166667	0	1.0000000	1.0000000	0.04830459
2	0.04666667	3	0.8400000	0.9733333	0.04792772
3	0.01833333	4	0.7933333	0.8433333	0.04582467
4	0.01400000	6	0.7566667	0.8333333	0.04564355
5	0.01333333	11	0.6866667	0.8300000	0.04558253
6	0.01000000	12	0.6733333	0.8600000	0.04612013

Models across the 50-50, 70-30, 80-20 training-test splits:

We made models based on three different partitions. As we can see 50-50 is the best model with highest accuracy.

50-50	Training	Validation
No seed	0.82	0.79
Seed (123)	0.83	0.81

70- 30	Training	Validation
No seed	0.80	0.79
Seed (123)	0.82	0.80

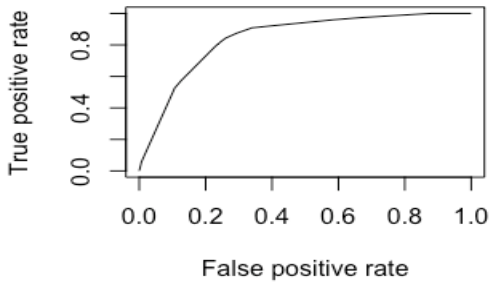
80-20	Training	Validation
No seed	0.81	0.78
Seed (123)	0.82	0.81

The split 50-50 is the best. Different number of seed is considered:

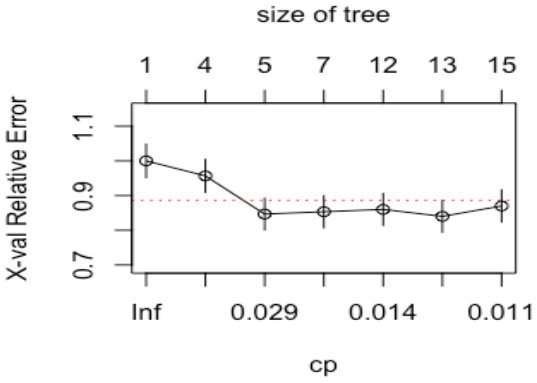
By looking at the below table, it is clear that with performance different sample of training and test data by using various seed. We found our model is almost is stable.

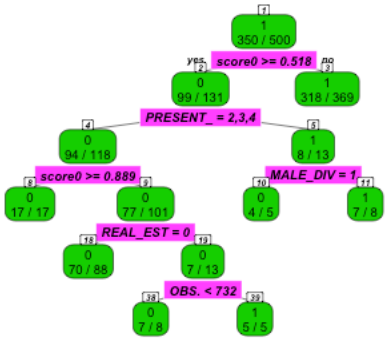
50-50	Training	Validation
No seed	0.82	0.79
Seed(123)	0.83	0.81
Seed(10)	0.81	0.80
Seed(100)	0.80	0.81
Seed(1000)	0.81	0.80

ROC:



plot of complexity parameter:





CTHRESH	ACCURACY OF TRAINING SETS
0.5	0.79
0.7	0.78
0.8	0.70

Decision tree C5.0

Confusion table:

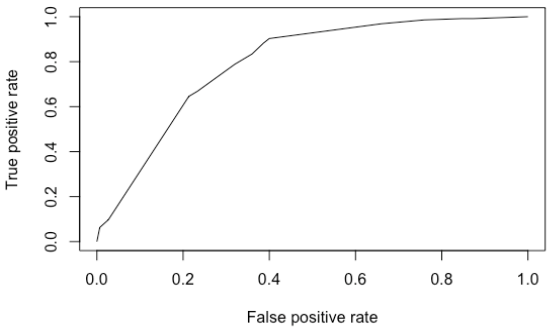
	0	**1**
0	237	23
1	63	677

The accuracy is: 0.91

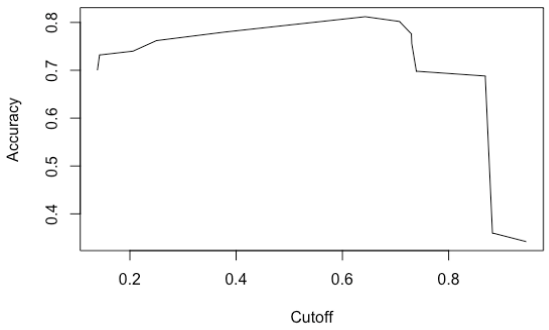
Classification Tree,
Number of samples: 1000
Number of predictors: 31
Tree size: 105
confidence level: 0.5

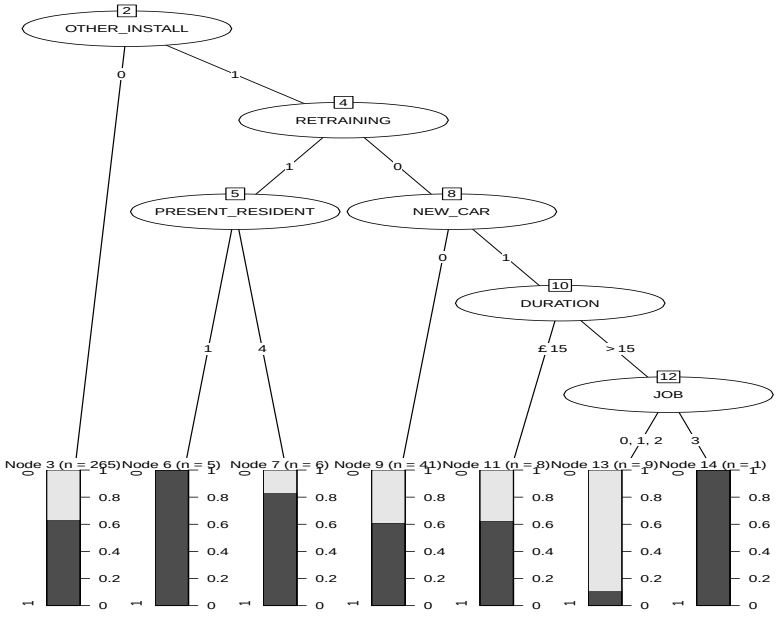
As we can see plot ROC shows, the better performance is at False positive rate four and True positive rate at none. When we have higher True positive rate and lower False positive rate we can have a better performance.

ROC:



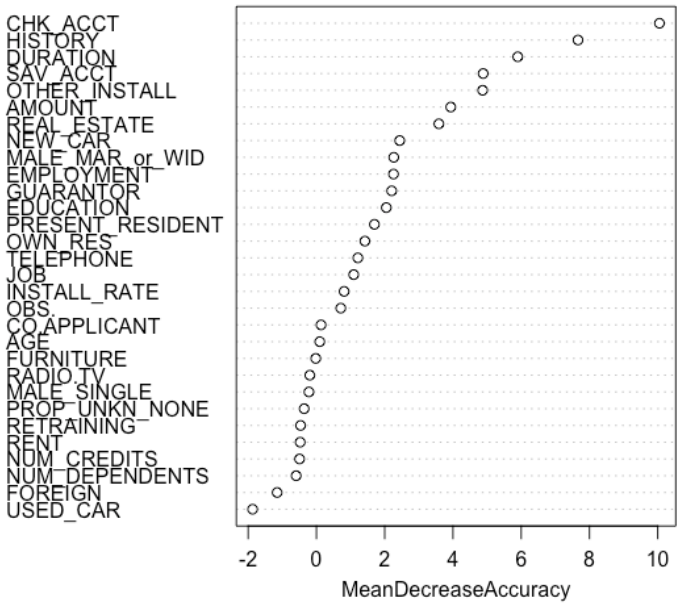
ACC:

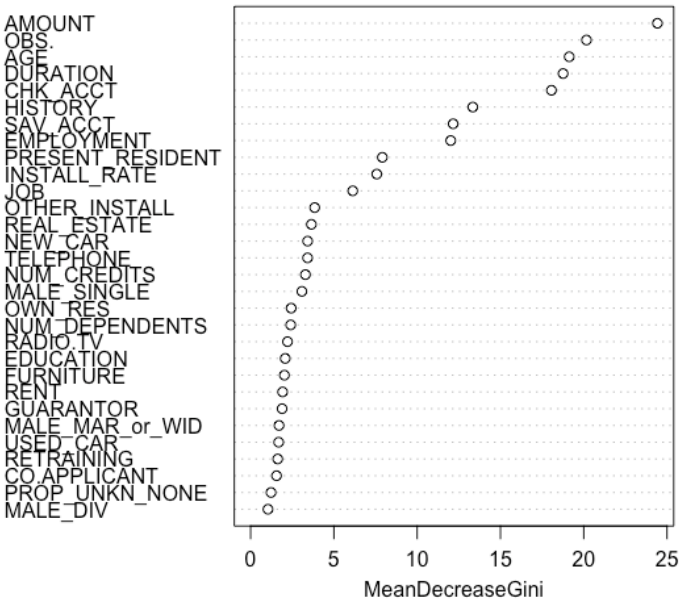




Random Forest:

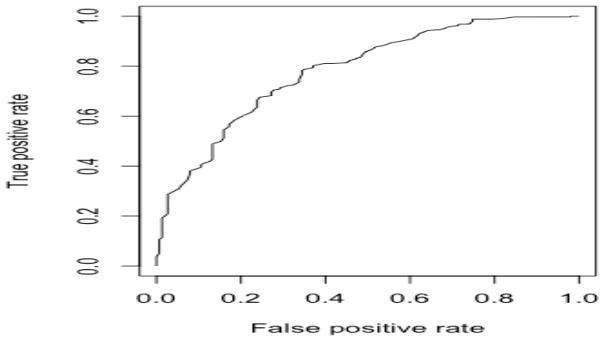
rfModel





The accuracy is : 1

ROC:



Confusion table:

	0	**1**
0	149	0
1	0	351

3. Consider the net profit (on average) of credit decisions as: Accept applicant decision for an Actual "Good" case: 100DM, and Accept applicant decision for an Actual "Bad" case: -500DM This information can be used to determine the following costs for misclassification:

		Predicted	
Actual		Good	Bad

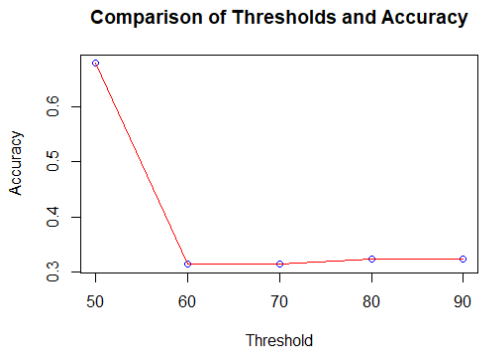
	Good	0	100 DM
	Bad	50	0

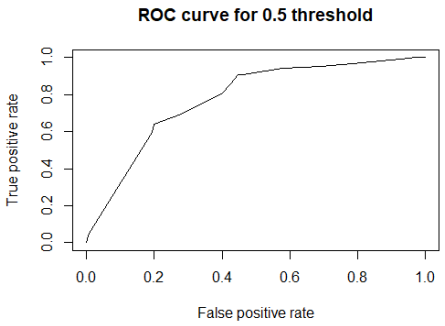
(a) For this analysis, we chose 70-30 model and incorporated in the cost sensitive matrix to form a decision tree (rptree73). The costs depend on the true and predicted class label. It shows an accuracy of 31.429%. To evaluate their performances, we considered the partitioned data for training and testing and checked their accuracy along with confusion matrix observations.

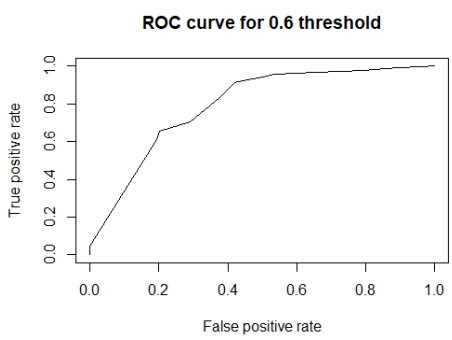
		True	
Pred		0	1
	0	202	480
	1	0	18

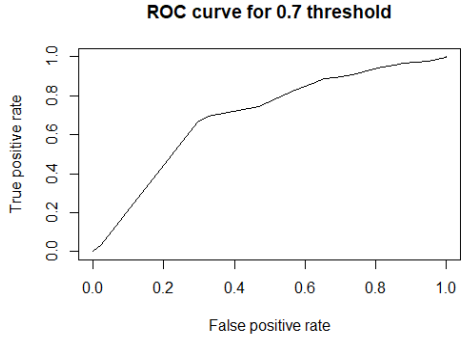
The accuracy was basically the mean of how much the training and tested data matched. Along with that we also check the ROC curves for performances. ROC curves incorporate the true positive and false positives rates, making our analysis better.
Different cutoff values were used for classification threshold. The thresholds used to turn posterior probabilities into class labels are chosen such that the costs are minimized. The analysis is tabulated and represented as follows:

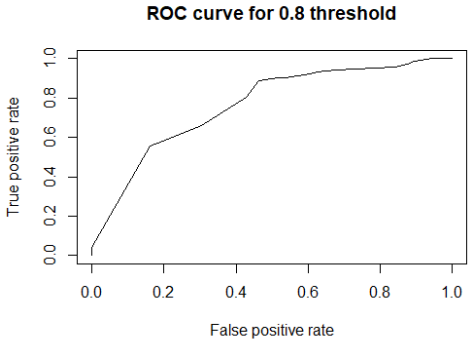
Threshold	Accuracy
0.5	0.6785714
0.6	0.3233333
0.7	0.3142857
0.8	0.325
0.9	0.3244444

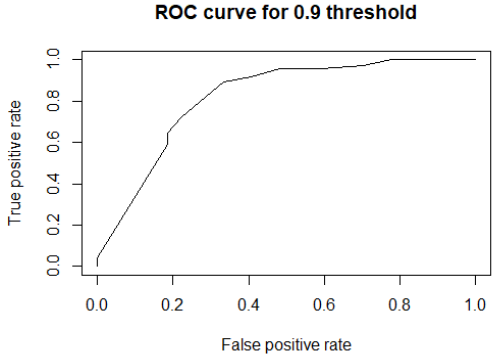












With changing threshold by 10%, there is a little change in accuracy. So, it can be seen from the ROC curve that the model chosen by us is giving the least accuracy. 80-20 is a little better than the chosen model. So, comparatively 80-20 model is giving best model.

(b) Calculate and apply the 'theoretical' threshold and assess performance – what do you notice, and how does this relate to the answer from (a) above.

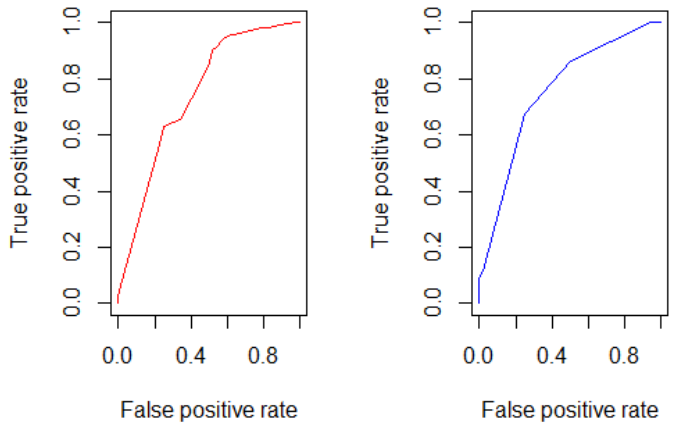
Theoretical threshold was calculated in RStudio as
th = costMatrix[2,1]/(costMatrix[2,1] + costMatrix[1,2]
The diagonal of the cost matrix is zero the formula given above simplifies accordingly.
This resulted in 0.833.
First we separated training and validation data then incorporated the threshold in 70-30 cost sensitive classification model. It has an accuracy of 31.573%.
In the second analysis, we defined a 83-17 cost sensitive model and incorporated it with the the theoretical threshold. It has an accuracy of 30.6122%.
To compare with the results of 3a, these accuracies are lower than the chosen model. So, we can say that the theoretical threshold won't give good performance if applied practically. The confusion tables and plots are as follows:

		True	
Pred		0	1
	0	247	565
	1	1	20

Confusion Table for First theoretical threshold analysis (in 70-30 model)

		True	
Pred		0	1
	0	245	575
	1	0	10

Confusion Table for Second theoretical threshold analysis (in 83-17 model)

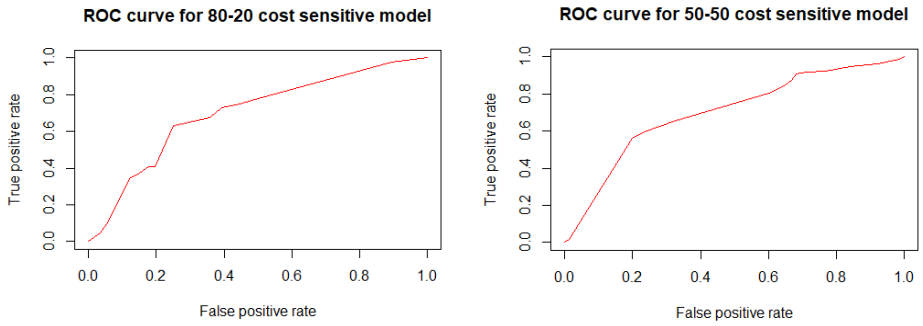


(c) We used the misclassification costs to develop the tree models (rpart and C5.0). Are the trees here different than ones obtained earlier?
Yes, the trees here are different than ones obtained earlier. They are more homogenized (pure), robust compared to earlier analyses. And have a well partitioned, organized and clearer data that can be used to calculate the credits.

Comparison performance of these two new models with those obtained earlier is as follows:
80-20 cost sensitive tree model gives an accuracy of 34.875%, which is little more than earlier observations.

		True	
Pred		0	1
	0	244	521
	1	0	35

Confusion Table for 80-20 cost sensitive model



50-50 cost sensitive tree model gives an accuracy of 81.4%.

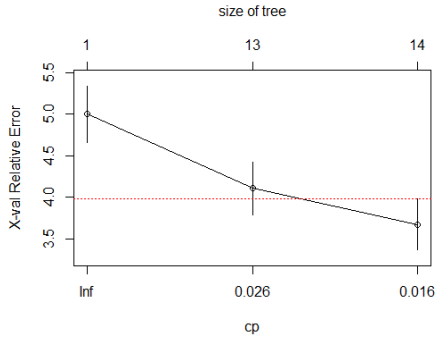
		True	
Pred		0	1
	0	94	37
	1	56	313

Confusion Table for 50-50 Cost Sensitive model

While other models improved just a little with around 5-6% maximum, the 50-50 model was default and has got significantly better than earlier analysis (more than 20%). So, this one could be used for further analysis.

4. The best decision tree model is the 50-50 cost sensitive model.

It has tree depth of 8.
It has 12 nodes.



The important variables for classifying good vs bad credit would include RESPONSE, CHK_ACT and AMOUNT. Relatively pure nodes are the nodes that are homogeneous or are close to being in a homogenized form. The two relatively pure nodes are:

1. IF(CHK_ACT =1 AND AMOUNT<3896 AND AMOUNT< 1991 AND DURATION<22 AND OBS<114
In the above case the probability of good case is 1.00 and bad case is 0.00

2. IF(CHK_ACT =0 AND AMOUNT>=3896 AND SAV_ACCT=1)
In the above case the probability of good is 0.82 and bad is 0.18.