IDS 572: Assignment 1

Decision Tree Analysis -- German Credit Data

Archan Patel - 661105271 - apate381@uic.edu

Jasbir Singh - 651837003 - jasingh3@uic.edu

Jeetyog Rangnekar - 656052696 - jrangn2@uic.edu

Table of Contents

Assignment Question 1

- Ratio of 'Good' to 'Bad' cases = 700/300=2.33

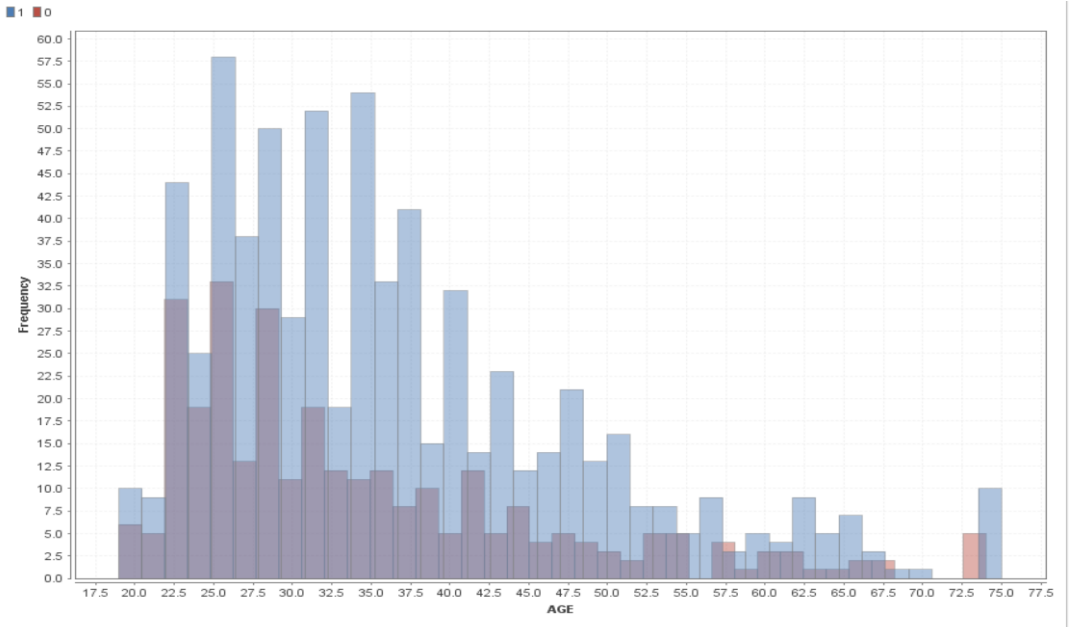
- Most of the missing values are observed in following attributes
  - NEW_CAR: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 76.6% of the data is missing
  - USED_CAR: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 89.7% of the data is missing
  - FURNITURE: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 81.9% of the data is missing
  - RADIO\TV: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 72% of the data is missing
  - EDUCATION: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 95% of the data is missing
  - RETRAINING: Binominal variable, where Yes=1, No= Missing data -- drop the variable since 90.3% of the data is missing
  - AGE: Integer variable: 9 missing data -- replace the missing data with the average of the dataset

- Description of Independent variables:

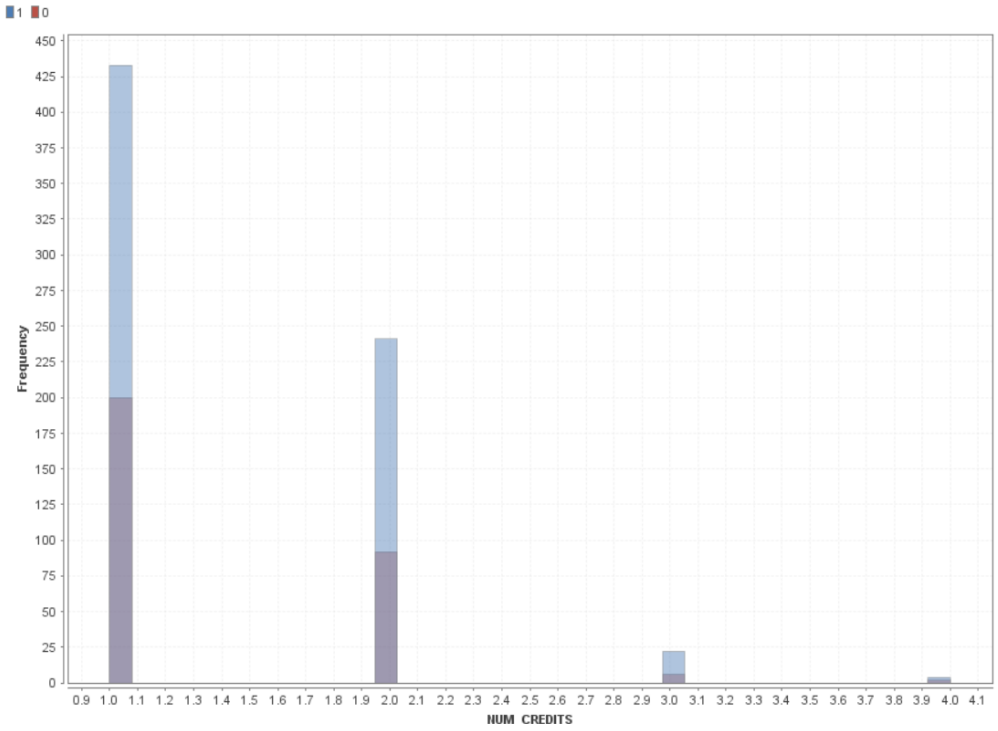
Variable Name	Type	Mean	Standard Deviation	Freq
CHK_ACCT	Categorical	--	--	0->274 1-> 269 2->63 3->394
DURATION	Integer	20.903	12.05	---
HISTORY	Categorical	--	--	0->40 1->49 2->530 3->88 4->293
AMOUNT	Integer	32751.156	2822.625	--
SAV_ACCOUNT	Categorical	--	--	0->603 1->103 2->63 3->48, 4->183
INSTALL_RATE	integer	2.973	1.119	--
EMPLOYMENT	Categorical	--	--	0->62 1->172 2->339 3->174 4->253
PRESENT_RESIDENT	Categorical	--	--	1->130 2->308 3->149 4->413
REAL_ESTATE	Categorical	--	--	0->718 1->282
AGE	integer	35.483	11.371	--
OTHER_INSTALL	Categorical	--	--	0->814 1->186

OWN_RES	Categorical			1->713 0->287
NUM_CREDITS	integer	1.407	0.578	--
JOB	Categorical	--	--	0->22 1->200 2->630 3->148
NUM_DEPENDENTS	integer	1.555	0.362	--
FOREIGN	Categorical	--	--	1->963 0->37

- Data Explore & Relationship
  - Customer of Age range from 20 till 30 are more likely to turn to bad creditor. This is little surprising.

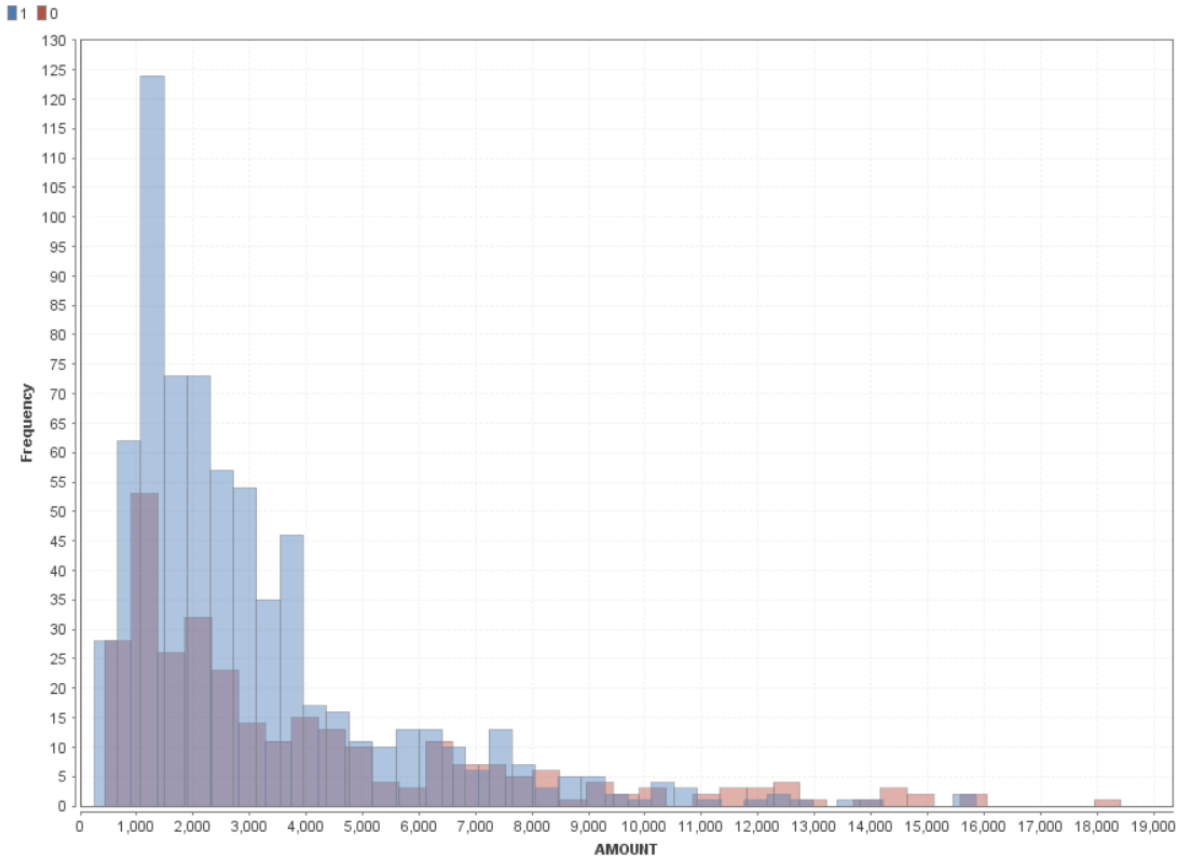


  - Lower Credit Score is bad creditor as expected .

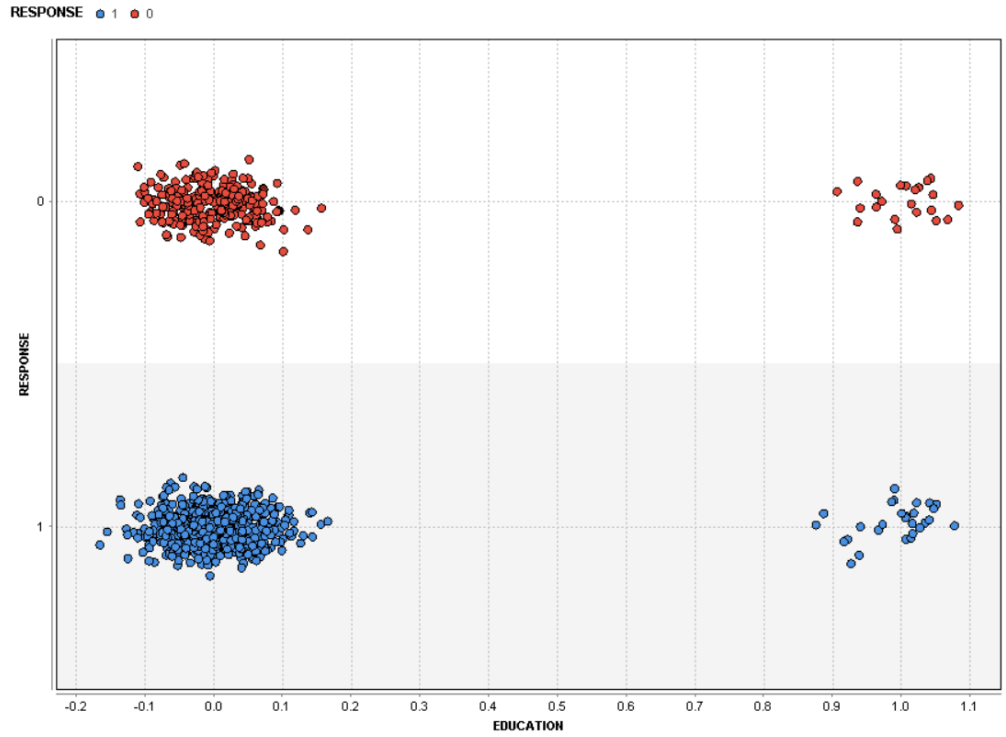


  - Interesting Relations:

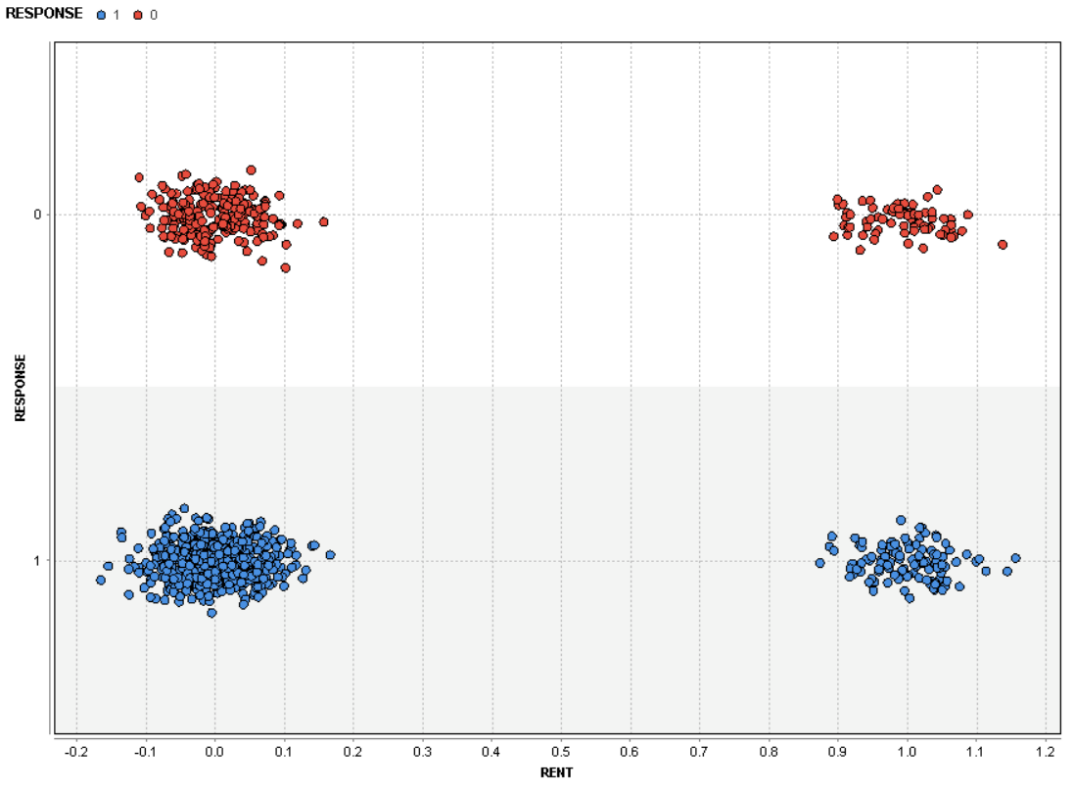
- Amount & Response : As amount increase ratio of BAD creditor increases



- Education & Response; Non-Educated customers are good creditor. It should be other way around



- RENT & RESPONSE: Customers that RENT are bad creditors



- Variables that will be most relevant for the Outcome:

Variable Name	Why ?
DURATION	Longer the duration, less like of BAD credit
HISTORY	Good credit history more possibility of GOOD Credit
AMOUNT	Less amount, less likelihood of BAD credit
SAV_ACCOUNT	More balance in the saving account, more likelihood of GOOD credit
INSTALL_RATE	More installment rate will lead to more BAD credits
EMPLOYMENT	Less duration of Employment, more chances of BAD credit
PRESENT_RESIDENT	More duration at resident, less likelihood of credit turning BAD
REAL_ESTATE	Applicant owns real state, less possibility of BAD credit
AGE	Older the age, less possibility of BAD credit
OTHER_INSTALL	Applicant has other credit plan, more possibility of BAD credit
OWN_RES	Applicant owns residence, less possibility of BAD credit
NUM_CREDITS	GOOD credit, less possibility of BAD credit
JOB	More skilled the Job, less possibility of BAD credit
NUM_DEPENDENTS	More number of dependent, more possibility of BAD loan
FOREIGN	Foreigner are less likely to default on the loan
CHK_ACCT	Checking account status should have direct impact on GOOD/BAD credit

Assignment Question 2

Full Data Decision Tree

(a)

- In regards to the full data, a good model of decision tree can be obtained using 'gain_ratio' criteria with maximal depth of 16.
  Decision tree with 'Gini_index' & 'Information_gain' are undesirable because they lead to small number of

records in each node - which becomes unstable model with unequal distribution. Among all criteria for splitting the tree, the 'gain_ratio' seems to be the only suitable criteria which does not give too many splits with small number of records in each node. There are fair amount of splits and records in each node which gives an idea of 'gain_ratio' to be a suitable criteria for splitting the decision tree.

(b)

- According to the Decision tree, following variables are important to differentiate "good" & "bad" credit cases.

AMOUNT	DURATION	CHK_ACCT	MALE_SINGLE	OTHER_INSTALL
DURATION	HISTORY	NUM_CREDITS	EMPLOYMENT	MALE_DIV

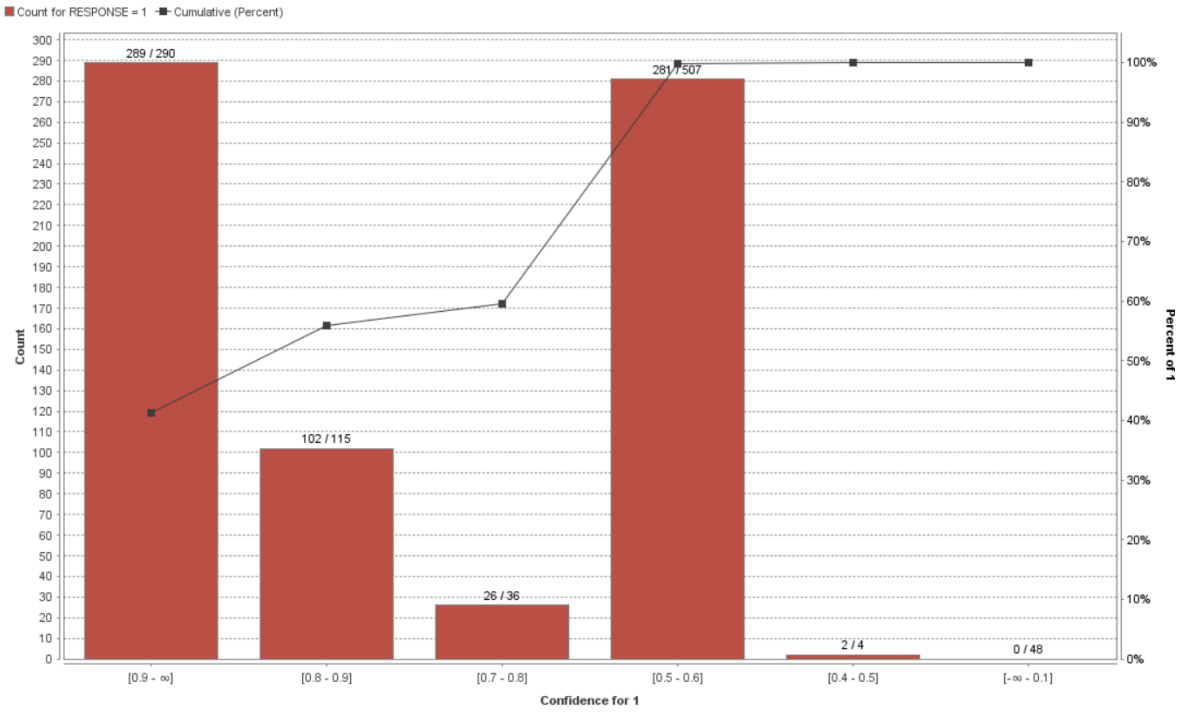
  - These variables are important because the decision tree has used these variables to classify the good and bad credit cases. And splits done on these variables at various stages of the tree provides a significant number of records in the leaf nodes with relatively high purity.
- Out of expected variables in Question 1, MALE_SINGLE & MALE_DIV are the variable which were not expected to be important ones.
  Other than that, SAV_ACCOUNT, INSTALL_RATE, PRESENT_RESIDENT, REAL_ESTATE, AGE, OWN_RES, NUM_DEPENDENTS & FOREIGN were expected to be important, but does not seem so.

(c)

- Below is the summary of performance parameters of different decision trees with 'gain_ratio' criteria.

Maximal Depth	Accuracy	Accuracy of "good"	Accuracy of "bad"
12	72.10%	71.50%	100.00%
14	73.20%	72.31%	100.00%
16	74.80%	73.53%	100.00%

- Lift Chart:



  - Within the confidence interval of 0.9 and above model is able to predict 269 out of 290 observations for good credit , Cumulative curve is increasing as confidence interval decreases

- This model cannot be considered a reliable model. Because it has not been tested yet. In order to declare a model reliable, it needs to be tested on a training data set.

Training & Testing Data Decision Tree

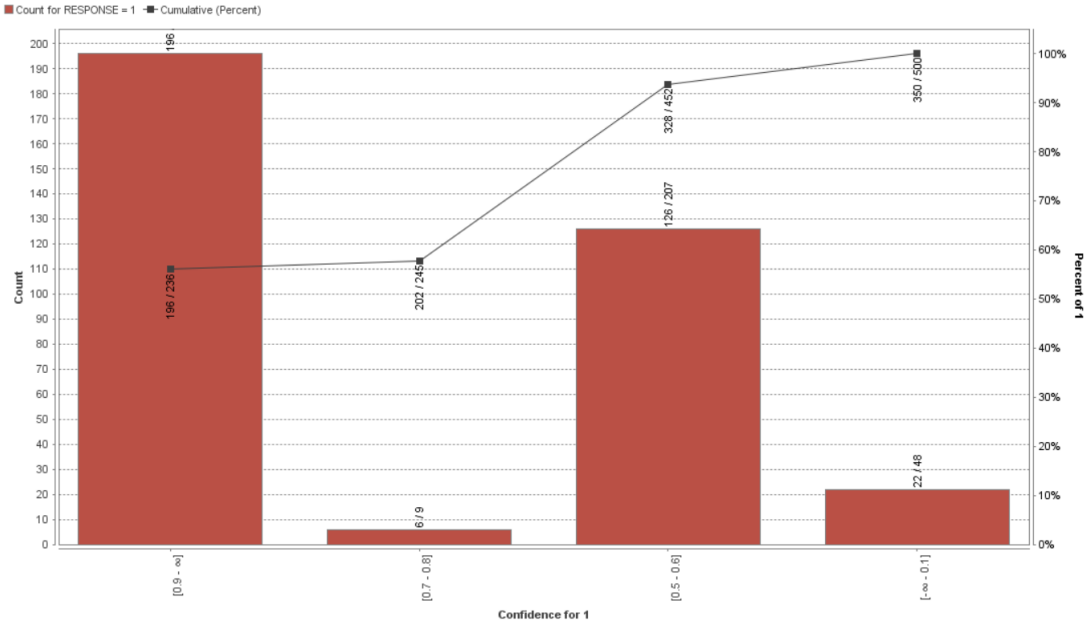
(a)

- The decision tree parameter 'Gini_index' & 'Information_gain' are giving almost 100% accuracy in training data & around 65% accuracy in testing data. This will result in overfit. 'Gain_ratio' seems to give fair amount of accuracy in testing & training dataset.
- Here is the summary of the Decision tree model obtained after splitting the data into 50-50 partition and with splitting criteria 'gain_ratio'.

Tree Depth	Accuracy		Precision		Recall		FP Rate	
	Training	Test	Training	Test	Training	Test	Training	Test
12	73.20	70.60	72.31	71.46	100.00	96.57	89.33	90.00
14	74.40	67.40	73.22	70.55	100.00	91.71	85.33	89.33
16	75.80	67.20	74.31	70.67	100.00	90.86	80.67	88.00

  - As displayed in the above table, the performance measures being used are Accuracy, Precision, Recall & FP Rate. The values for these measures obtained for difference set of maximum tree depth values have been observed in the table. The reason of using this measures is, overall they give a better picture about the performance of the model.
  - The higher the accuracy of the training dataset, higher the chances of overfit. So, accuracy has been considered here.
  - With higher precision, we can get the probability of predicting the true positive of the model which is very essential in predicting 'good' or 'bad' credit.
  - Higher recall with a comparatively lower FP rate leads to good performance of the model.
  - With FP Rate, we can determine the ability of predicting 'bad' credit of the model.

- Lift Chart:



- This model still cannot be considered as a reliable model, because the model does not have the enough dataset for the training & the FP rate of the model is high - which means model is unable to predict the 'bad' credit precisely.
- The pruning does not give a better model, because it does not significantly impact any of our performance parameters
- So as explained above, for developing a good model, we can consider Accuracy, Precision, Recall as well as FP Rate & Lift to be useful since they give a clear overall picture of the model. They differ for different training-test data partitions. Below are a list of their values for few of training-test data partition samples.

Tree Depth		**50-50**		**70-30**		**80-20**	
	Measure	Training	Test	Training	Test	Training	Test
12	Accuracy	73.20	70.60	74	67.67	73.12	70
	Precision	72.31	71.46	72.99	69.82	72.26	70.62
	Recall	100	96.57	99.8	94.76	100	97.86
	FP Rate	89.33	90.00	86.19	95.5	89.58	95.00
14	Accuracy	74.4	67.4	74.43	68.00	73.25	70
	Precision	73.22	70.55	73.24	69.93	72.35	70.62
	Recall	100	91.71	100	95.24	100	97.86
	FP Rate	85.33	89.33	85.23	95.5	89.16	95.00
16	Accuracy	75.8	67.2	74.86	67.67	**74.38**	**68**
	Precision	74.31	70.67	73.57	69.82	**73.26**	**70.21**
	Recall	100	90.86	100	94.76	**99.82**	**94.29**
	FP Rate	80.67	88.00	83.81	95.5	**85**	**93.33**

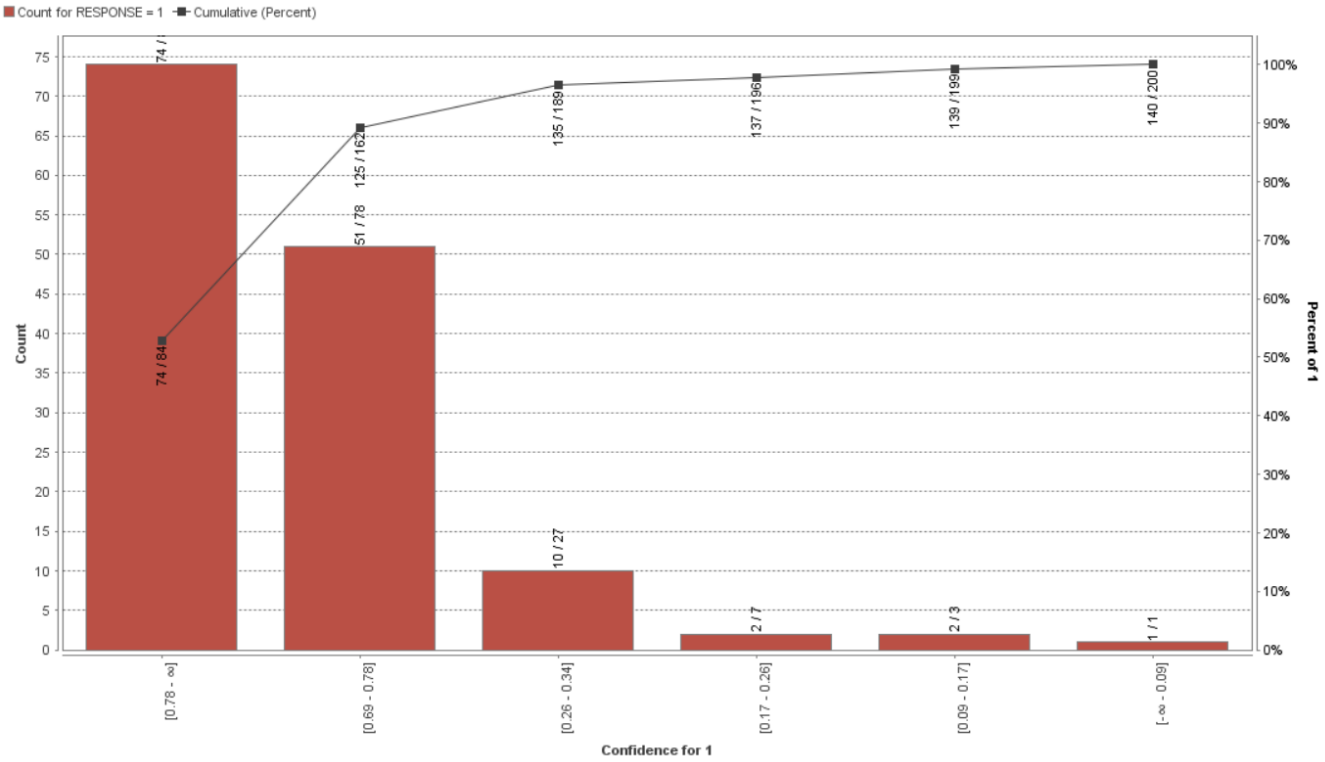
(b)

- The other decision tree operators considered: CART & J-48
- Below is the summary of CART operator.
    - S = Random Number Seed
    - M = Minimum number of instances at terminal node
    - N = The number of folds used in the minimal cost-complexity

Parameter		**50-50**			**70-30**		**80-20**	
	Measure	Training	Test		Training	Test	Training	Test
		70-30						

S=2 M=4 N=5	Accuracy	*84.43*	*72.67*	**S=1 M=2 N=10**	82.29	73	81.25	70.50
	Precision	*86.42*	*77.35*		83.89	77.22	81.73	74.85
	Recall	*92.24*	*86.19*		92.45	87.14	94.29	87.14
	FP Rate	*33.8*	*58.8*		41.4	60	49.16	68.33
S=3 M=4 N=5	Accuracy	80	71.60		78.14	75.33	**78.62**	**74**
	Precision	81.73	77.51		78.61	76.36	**79.97**	**77.16**
	Recall	92	83.71		94.49	93.81	**92.68**	**89.29**
	FP Rate	48	56.67		60	67.7	**54.16**	**61.67**
S=7 M=2 N=5	Accuracy	77.8	69.8		78.14	75.33	80.75	72
	Precision	79.65	75.71		78.61	76.3	82.85	76.92
	Recall	91.71	83.71		94.49	93.81	91.43	85.71
	FP Rate	54.67	62.67		60	67.7	44.16	60

- Lift for CART operator with 80:20 partition in dataset & S=3,M=4, N=5

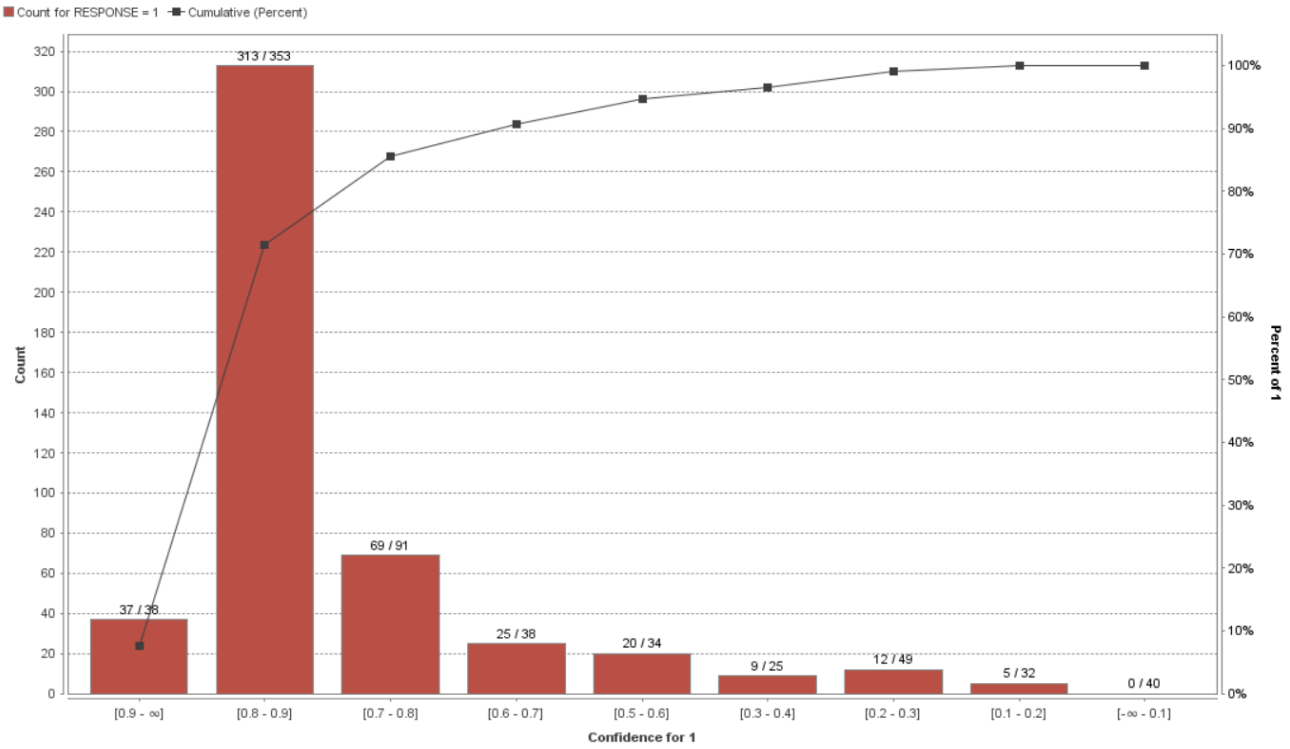


  - In the confidence interval of 0.9 and above, there are 88% of true positive cases indicating good credit risks and 65% good credit in the 0.8-0.9 confidence which is less than the Naive case.

- Below is the summary of J48 Operator.
    - C = Confidence threshold for pruning
    - M = Minimum number of instances per leaf
    - N = Number of folds for reduced error pruning

Parameter		**50-50**		**70-30**		**80-20**	
	Measure	Training	Test	Training	Test	Training	Test
C=.25 M=2 N=3	Accuracy	79.8	73	80.29	71	79.12	71
	Precision	81.89	78.67	82.84	75.95	82.06	79.29
	Recall	91.43	84.29	90.61	85.71	89.82	79.29
	FP Rate	47.33	53.33	43.8	63.33	45.83	48.33
C=.25 M=3 N=6	Accuracy	81.40	69.60	79.57	73	77.62	71.50
	Precision	87.03	78.78	81.37	77.45	76.57	72.43
	Recall	86.29	77.43	91.84	86.67	98.04	95.71
	FP Rate	30	48.67	49.04	58.9	70	85
C=.5 M=6	Accuracy	84.20	72.23	**83**	**74.33**	83.30	72
	Precision	86.33	78.59	**84.29**	**79.56**	85.91	78
	Recall	92	82.86	**93.06**	**85.24**	91.43	83.57
	FP Rate	34	52.67	**40.4**	**51.11**	35	55

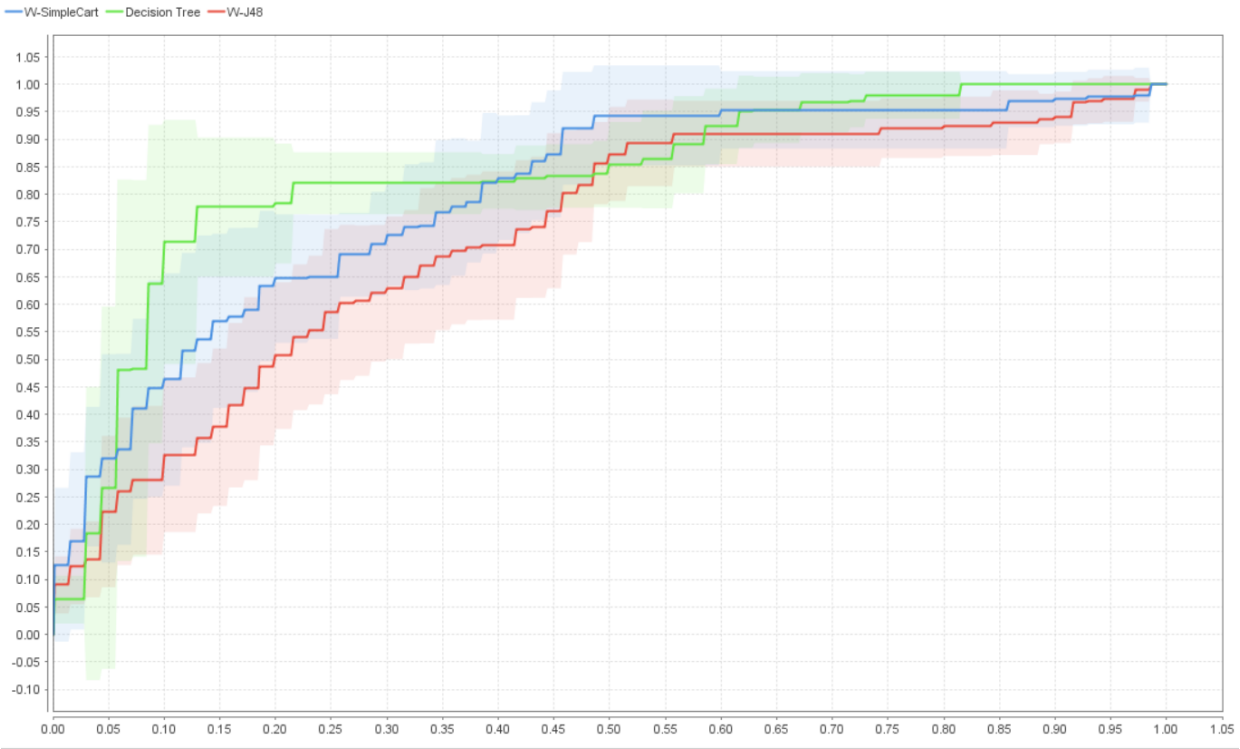
- Lift of J48 operator with 70:30 partition in dataset & C=0.5, M=6



- In the confidence interval of 0.9 and above, there are 97% of true positive cases indicating good credit risks and 88.6% good credit in the 0.8-0.9 confidence which is better than the CART algorithm.

- From the summary tables & Lift, it is clear that performance differ across different decision tree learners as you change the parameter values. For the comparison purpose, same performance measures (Accuracy, Precision, Recall & FP Rate) has been considered.

- Here is the ROC curve comparison for three operators.



- From the ROC curve, it can be seen that Decision Tree model is better among three, while CART maybe little better in some of the parts. J48 is worse among the three.

(c)

- Here is the summary of Decision tree with different random seed values on 70-30 partition of data..

Random Seed	**10**		**123**		**1992**		**9999**	
Measure	Training	Test	Training	Test	Training	Test	Training	Test
Accuracy	76.71	64.67	75.57	68.67	78.29	69	74.43	67.67
Precision	75.12	69.26	74.2	70.42	74.35	71.12	73.24	69.96
Recall	99.80	89.05	99.80	95.24	98.78	93.81	100	94.29
FP Rate	77.14	92.2	80.9	93.3	79.5	88.89	85.2	94.44

  - It can be observed that changing the random seed value does not affect the performance of the model. So, we can conclude that the model is stable.

(d)

- Considering the data mentioned in the above tables, here is the comparison between the 'good' model performance of each decision tree learner.

Decision Tree Learner & Parameter	**Decision Tree Partition: 80-20 Tree Depth: 16**		**CART Partition: 80-20 S=3, M=4, N=5**		**J48 Partition: 70-30 C=0.5, M=6**	
Measure	Training	Test	Training	Test	Training	Test
Accuracy	74.38	68	78.62	74	83	74.33
Precision	73.26	70.21	79.97	77.16	84.29	79.56
Recall	99.82	94.29	92.68	89.29	93.06	85.24
FP Rate	85	93.33	54.16	61.67	40.4	51.11

  - In earlier questions, we had considered Accuracy, Precision, Recall, FP Rate & Lift to be the parameters to get a good model. Those parameters seems to be giving good overview here as well.

- The variables like Amount, Duration, Chk_Acct are used consistently by all model in the upper part of the tree.
  - Amount, Duration and Chk_Acct are the most decisive variables with a higher split ratio which have a major influence on the Response variable for classification of good and bad cases.
  - We choose the Decision Tree with a maximum depth of 16 and a training-testing split ratio of 80:20.

Question 3

- We make use of the following credit decisions to calculate the net profit of our model:
  - Accept applicant decision for an Actual "Good" case: 100DM, and
  - Accept applicant decision for an Actual "Bad" case: -500DM

- From the above credit decisions we have the cost decision matrix as follows:

	Predicted 1	**Predicted 0**
Actual 1	100 DM	0
Actual 0	-500 DM	0

- From the above matrix, we understand that, for every customer correctly predicted as a good credit we have a cost benefit of 100 DM and for every customer incorrectly predicted as a good credit the bank loses 500 DM.
From our model, we obtain the following confusion matrix for the Validation data:

	Predicted 1	**Predicted 0**
Actual 1	132	8
Actual 0	56	4

- From the above table we can now calculate the overall cost benefit for validation data as:

Net profit = 132 x 100 + 56 x (-500)

= - 14,800 DM

  - So, we see that the model gives us a net loss at the confidence threshold of 0.5. We calculate the Opportunity cost of lost credits using the below Cost Matrix:

	Predicted 1	**Predicted 0**
Actual 1	0	100 DM
Actual 0	500 DM	0

- The Opportunity cost of lost credits is given by the loss of 100 DM per customer who was actually a good credit but predicted as a bad credit risk and a 500 DM loss per customer who were bad credit risks but incorrectly predicted as good credits.

Opportunity costs of lost credits = 8 x 100 + 56 x 500

= 28,800

- As the confidence threshold of 0.5, gives us a net loss, we try changing the thresholds to get the maximum possible net benefit. Trying different levels of threshold lead us to the following findings:

Threshold (T)	**Misclassification Costs**
0.3	11300
0.4	11300
0.45	11500
0.5	28800
0.6	28800
0.7	28700
0.8	28700

As we can see from the above empirical findings that a confidence threshold of 0.4 gives us the least misclassification cost of 11300 DM. Thus, we can set the threshold to 0.4 to get the maximum net benefit.

Question 4

(a)

- The best model obtained has the tree depth of 16 with dataset partitioned in 80-20 ratio It has 64 nodes.
- The variables towards the top are AMOUNT, DURATION, CHK_ACCT & AGE.
    - All these variables except AGE was considered as Important variable in Question 2.

(b)

- Two relatively pure leaf
    - First one can be found at the split with the regards to the variable GUARANTOR which has 102 'Good' & 13 'Bad' cases.
        - Probability of 'Good': 88.7%
        - Probability of 'Bad': 11.3%
    - Second one can be found at depth with regards to the variable MALE_DIV which has 26 'Good' & 10 'Bad' cases.
        - Probability of 'Good': 72.22%
        - Probability of 'Bad': 27.78%

(c)

- Sample Rule 1:

IF the Credit Amount is <= 15901 AND
IF the Duration of credit in months is <= 66 AND
IF Checking Account status is >= 200 DM or 'no checking account'
IF Age is > 19.5 AND
IF Amount is > 10924 AND
IF Applicant is Single Male AND
IF Duration of credit in months is > 28.5
THEN Applicant is 'Good'

- Sample Rule 2:

IF the Credit Amount is <= 15901 AND
IF the Duration of credit in months is <= 66 AND
IF Checking Account status is <0 or between 0 to 200 AND
IF Credit Amount is > 645.5 AND
IF Age is > 74.5
THEN Applicant is 'Good'.

Question 5

- We retrieved the ExampleSet result in the attached Excel sheet (Question 5.xlsx) using a threshold of 0.4 as the misclassification cost in this case was minimum.
- We considered only, the "good" cases i.e Response = 1 as we have to interpret the good credit risks.

- After sorting the Confidence(1) in descending order and calculating the cumulative cost/net benefit we observed that we have a maximum profit of **3100** at a confidence of **93.42%** for "good" credit risk.

Performance Measures	Values
Accuracy	59.50%
Precision	89.33%
Recall	47.86%
FP Rate	13.33%

Random Forest

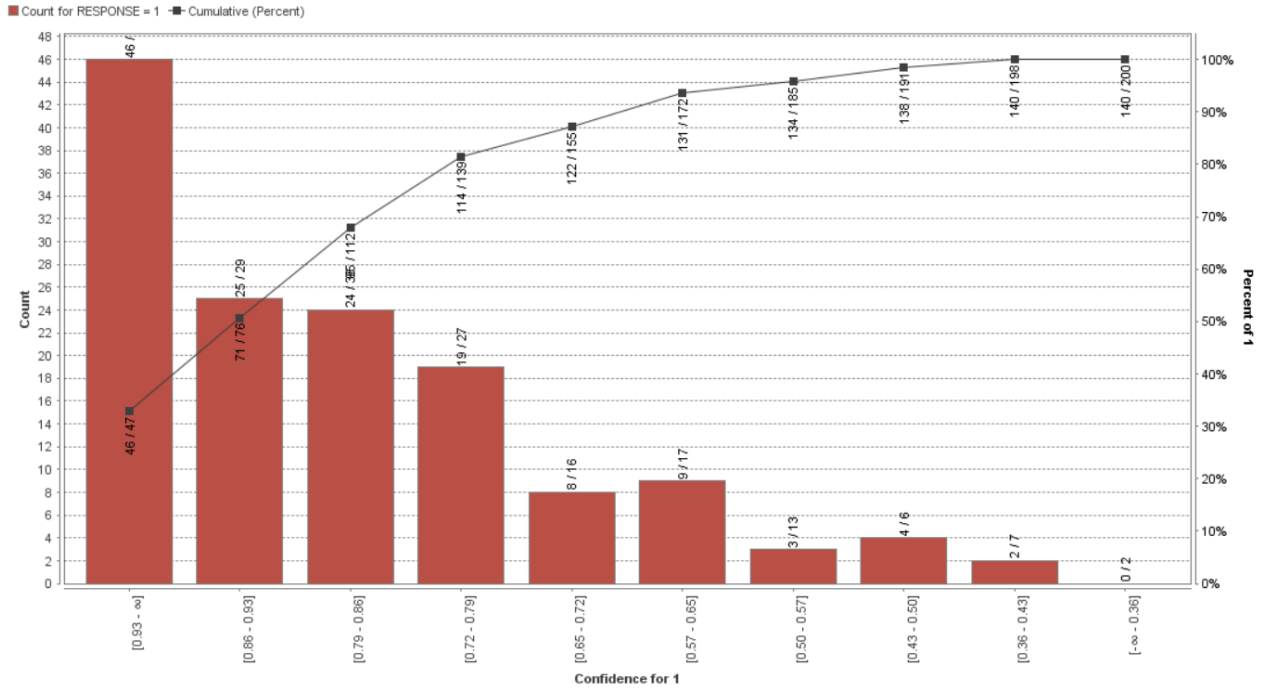
- Here is the summary of different random forest models with different parameters. It has been observed that best values can be found with subset ratio 0.6. While using the voting strategy 'confidence vote', the performance measure values were found to be way lesser than 'majority vote' - thus, 'majority vote' has been considered as voting strategy for below findings.

Parameter		**70-30**						**80-20**					
		N=100, D=16		N=20, D=12		N=200, D=14		N=100, D=16		N=20, D=12		N=200, D=14	
	Measure	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te	Tr	Te
GINI	Accuracy	82.43	71.33	81.86	73.33	81.86	72	83	72.5	79.25	72	82	72.5
	Precision	79.93	72.63	80.61	74.81	79.42	72.83	80.46	73.22	77.36	72.58	79.55	72.97
	Recall	100	94.76	97.55	93.33	100	95.71	100	95.71	99.46	96.43	100	96.43
	FP Rate	58.57	83.33	54.76	71.11	60.47	83.33	56.67	81.67	67.91	85	60	83.33
INFoGAIn	Accuracy	81	72.33	81.29	73	80.43	71	**81.88**	**75**	79.38	72.50	80.62	72
	Precision	78.65	72.92	79.77	74.52	78.15	72.04	**79.43**	**75**	77.55	73.22	78.32	72.83
	Recall	100	96.19	98.16	93.33	100	95.71	**100**	**96.43**	99.29	95.71	100	95.71
	FP Rate	63.33	83.33	58.09	74.44	65.23	86.67	**59.16**	**75**	67.08	81.67	64.58	83.33
GAInRaTiO	Accuracy	70.57	70.67	71	70.33	70.43	70	70.50	70	70.62	70	70.50	70
	Precision	70.40	70.61	70.71	70.37	70.30	70.13	70.35	70.20	70.44	70.20	70.35	70.20
	Recall	100	99.52	100	99.52	100	99.52	100	99.29	100	99.29	100	99.29
	FP Rate	98.09	96.67	96.67	97.78	98.57	98.89	98.33	98.33	97.92	93.33	98.33	98.33

- Changing the voting strategy to 'Confidence vote', the performance measure values were found to be lesser
- Considering the data above, the good model can be obtained with data partition 80-20, criteria 'Information_gain', number of trees =100 & tree depth=16. Using this model, summary of good cases to bad cases with different confidence intervals is listed below.

Confidence Interval	0.93-1	0.86-0.93	0.79-0.86
Ratio (good/bad)	0.97	0.93	0.85

- Lift chart:



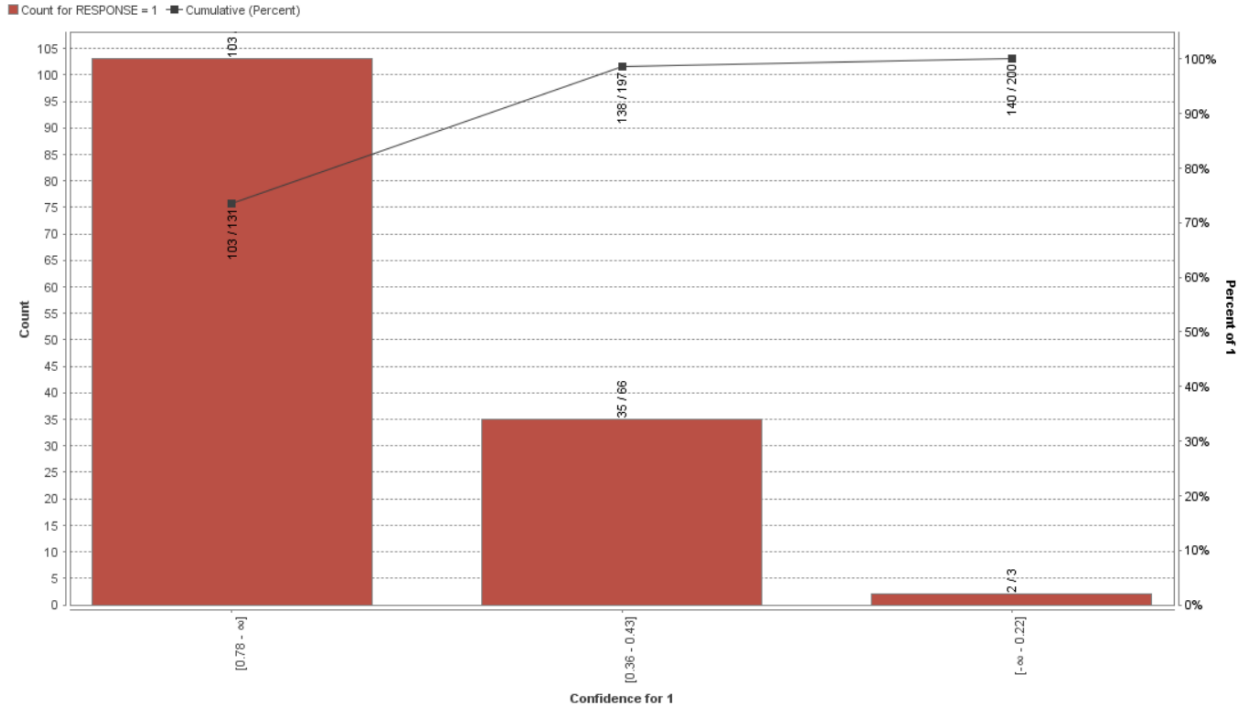
ADABoost

- Here is the summary of different ADABoost models with different parameters. It has been observed that a good model of decision tree can be obtained using 'gain_ratio' criteria. The other criteria provides an accuracy of 100% on the training data, which will lead to overfitting.
- Also, it is observed that number of iteration of ADABoost does not change the performance measures.

Tree Depth		**80-20**		**70-30**	
	Measure	Training	Test	Training	
12	Accuracy	**73.38**	**70**	73.14	70.
	Precision	**80.17**	**79.85**	72.27	70.
	Recall	**82.32**	**76.43**	100	98.
	FP Rate	**47.5**	**45**	89.52	94.
16	Accuracy	75.62	67.50	75	70.

	Precision	82.07	78.63	74.42	71.
	Recall	83.39	73.57	97.96	95.
	FP Rate	42.5	45	78.57	87.

- The good model can be found with tree depth of 12 and 80-20 dataset partition. The ratio of good to bad cases with confidence 0.76-1 is 0.79.
- Lift Chart:



Below is the ROC curve for the above mentioned good models of Random forest & ADABoost. From the curve, it is clear that ADABoost is a better model than Random Forest since AUC is larger for ADABoost model.