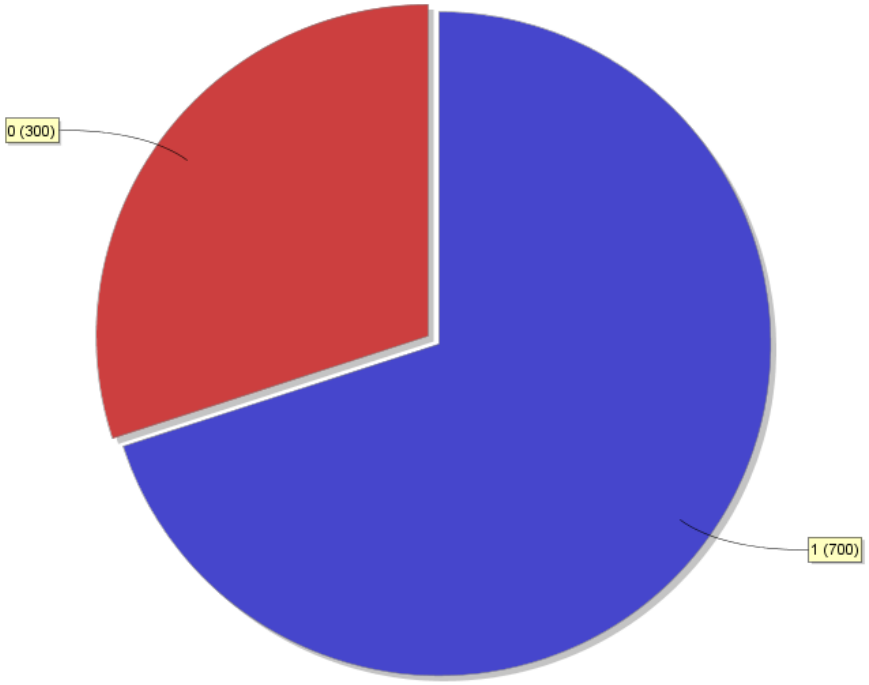
IDS 572
Assignment – 1
German Credit Data – Decision Trees

Bala Rohit Yeruva

Fall 2016

Answer #1

The proportion of 'Good' and 'Bad' cases in the given dataset are 700 (70%) and 300 (30%) respectively.



There are 30 independent variables with three different data types: Real values(Quantitative), Categorical and Logical.

Quantitative Variables: Statistics

Variable	'Good' Mean	'Bad' Mean	Total Mean	'Good' Standard Deviation	'Bad' Standard Deviation	Total Standard Deviation
Duration	19.207	24.86	20.90	11.08	13.28	12.06
Amount	2985.46	3938.13	3271.26	2401.47	3535.82	2822.74
Install_Rate	2.92	3.10	2.97	1.13	1.09	1.12
Age	36.22	33.96	35.55	11.38	11.22	11.38
Num_Credits	1.42	1.37	1.41	0.58	0.56	0.58
Num_Dependents	1.16	1.15	1.16	0.36	0.36	0.36

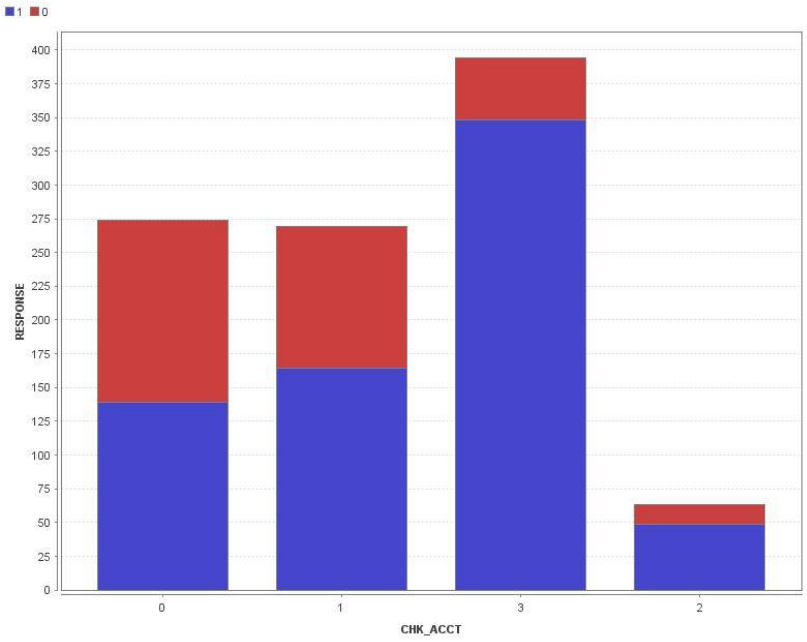
Note: 'Good' Mean and 'Good' Standard Deviation indicates the mean and standard deviation of the respective variable when the response is 'Good'. The same goes with 'Bad' Mean and 'Bad' Standard Deviation.

Logical Variables: Frequencies

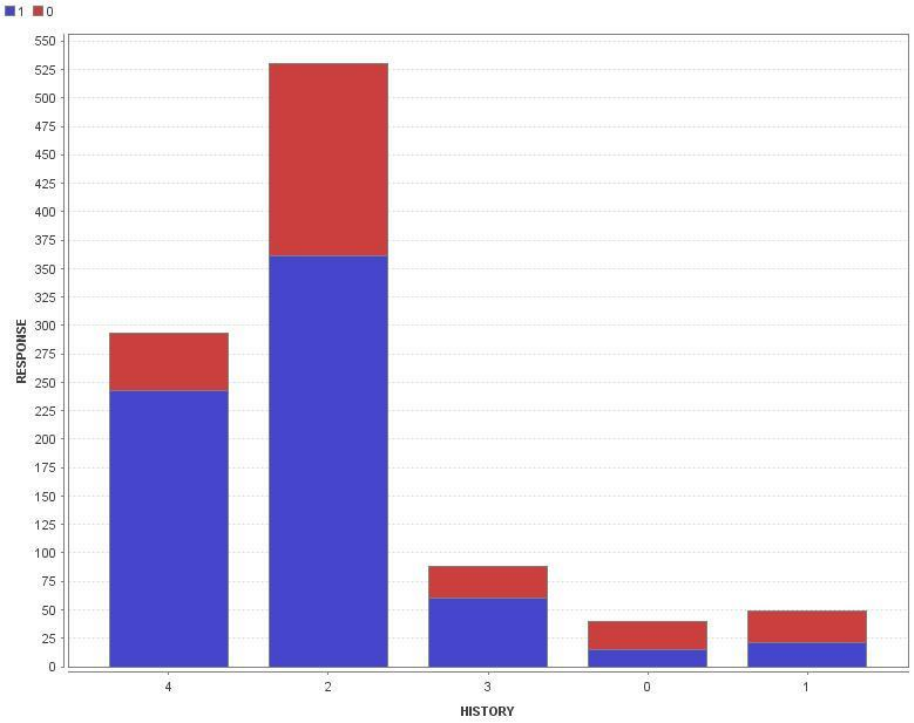
Variable	'Good' True	'Bad' True	'Good' False	'Bad' False
New_Car	145	89	555	211
Used_Car	86	17	614	283
Furniture	123	58	577	242
Radio/TV	218	62	482	238
Education	28	22	672	278
Retraining	63	34	637	266
Male_Div	30	20	670	280
Male_Single	402	146	298	154
Male_Mar_Wid	67	25	633	275
Co-Applicant	23	18	677	282
Guarantor	42	10	658	290
Real_Estate	222	60	478	240
Prop_Unkn_None	87	67	613	233
Other_Install	110	76	590	224
Rent	109	70	591	230
Own_Res	527	186	173	114
Telephone	291	113	409	187
Foreign	33	4	667	296

Categorical Variables: In all the graphs below, 'Good' is represented as blue and 'Bad' as red.

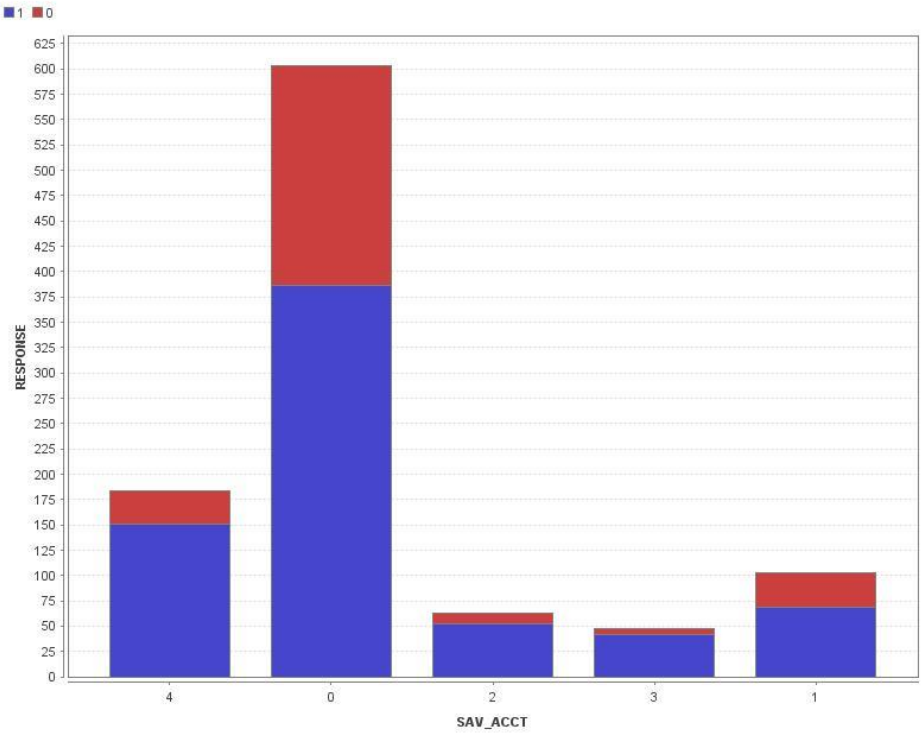
Variable: Checking Account (See Appendix A)



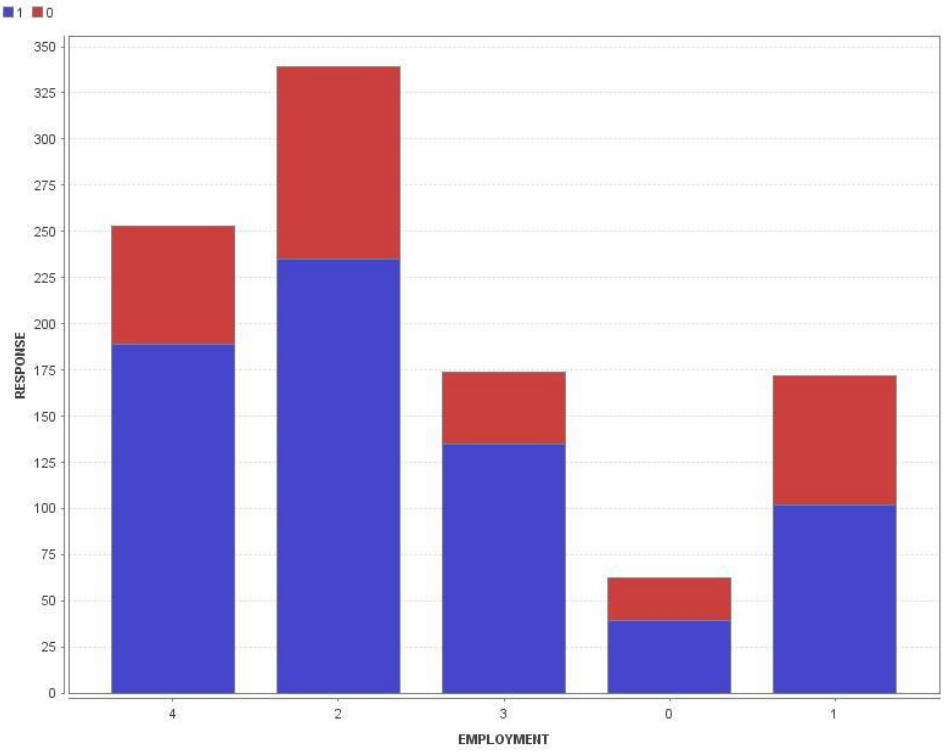
Variable: History (See Appendix A)



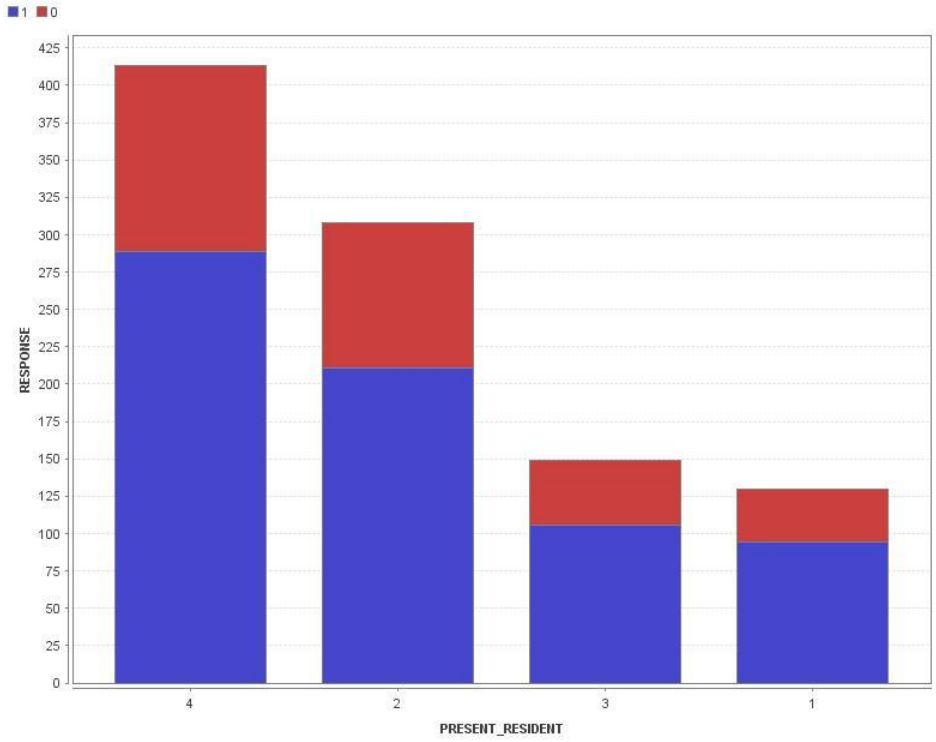
Variable: Savings Account (See Appendix A)



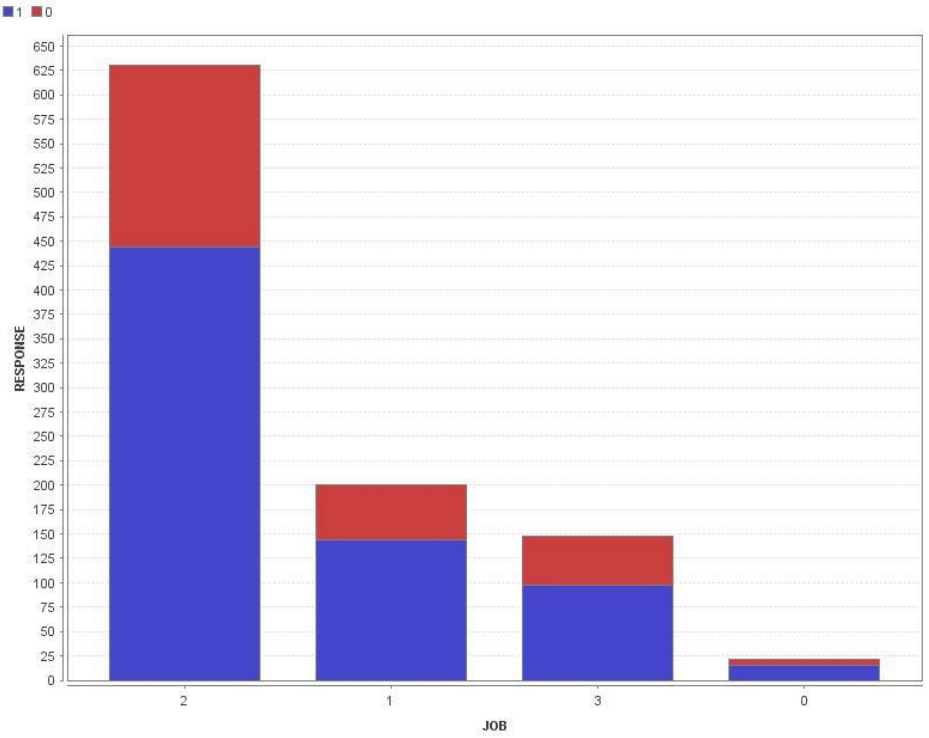
Variable: Employment (See Appendix A)



Variable: Present Resident (See Appendix A)



Variable: Job (See Appendix A)



From the descriptions and the distributions of the variables, we can infer that the following variables could be most influential in the classification of the Response variable: Checking Account, Duration, History, Amount and Saving Account.

Answer #2

A decision tree was built (See Appendix B) using the full data and the following variables were used to distinguish the 'Good' and 'Bad' cases:

- Checking Account Status (Chk_Acc)
- Credit History (History)
- Credit Amount (Amount)
- Duration of Credit in Months (Duration)
- Installment Rates as a % of disposable income (Install_Rate)
- If the applicant is a single male (Male_Single)
- If the applicant owns a residence (Own_Res)
- If the applicant owns a real estate (Real_Estate)
- If the applicant owns no property or unknown (Prop_Unkn_None)
- If the applicant has other installment plan credit (Other_Install)
- If the purpose of the credit is for a new car (New_Car)

This model had an accuracy level of 75.50% with the accuracy for 'Good' and 'Bad' cases being 79.43% and 62.11% respectively (See Appendix C).

A model defined as such is not reliable because we have used the entire data to build the tree without saving any data for validating the model. Hence, such a model has high chances of performing badly against any unseen data.

A few of the key parameters used for building a good model are Split Criteria, Maximal depth, prepruning, minimal gain, minimal leaf size. Split criterion is used to select the attributes for splitting. Maximal Depth is used to restrict the size of the decision tree. Prepruning is used to decide whether to split the node further based on certain parameters such as Minimal Gain and Minimal leaf size.

The following are the corresponding values assigned to the above parameters while building the descriptive model:

- Split criterion: Gini_Index
- Maximal Depth: 20
- Prepruning: Yes
- Minimal Gain: 0.07
- Minimal Leaf Size: 5
- Minimal Size for Split: 5
- Number of prepruning alternatives: 5

Answer #3

a. Partitioning of data into 50% of Training data and 50% of Validation data:

Performance on Training Data:
Accuracy: **79.80%**

	True 1	*True 0*	*Class Precision*
Pred. 1	305	51	85.67%
Pred. 0	50	94	65.28%
Class recall	85.92%	64.83%	

Performance on Validation Data:
Accuracy: **68.20%**

	True 1	*True 0*	*Class Precision*
Pred. 1	269	83	76.42%
Pred. 0	76	72	48.65%
Class recall	77.97%	46.45%	

Based on the accuracy levels, we can conclude that the training data was not sufficient enough to build the model which explains the low performance on the validation data. Therefore, this model is not considered to be robust.

b. Partitioning of data into 70% of Training data and 30% of Validation data:

Performance on Training Data:
Accuracy: **83.00%**

	True 1	*True 0*	*Class Precision*
Pred. 1	479	91	84.04%
Pred. 0	28	102	78.46%
Class recall	94.48%	52.85%	

Performance on Validation Data:
Accuracy: **69.67%**

	True 1	*True 0*	*Class Precision*
Pred. 1	174	72	70.73%
Pred. 0	19	35	64.81%
Class recall	90.16%	32.71%	

- Partitioning of data into 80% of Training data and 20% of Validation data:

Performance on Training Data:
Accuracy: **83.25%**

	True 1	*True 0*	*Class Precision*
Pred. 1	526	93	84.98%
Pred. 0	41	140	77.35%
Class recall	92.77%	60.09%	

Performance on Validation Data:
Accuracy: **72.50%**

	True 1	*True 0*	*Class Precision*
Pred. 1	116	38	75.32%
Pred. 0	17	29	63.04%
Class recall	87.22%	43.28%	

- Partitioning of data into 75% of Training data and 25% of Validation data:

  Performance on Training Data:
  Accuracy: **80.13%**

	True 1	*True 0*	*Class Precision*
Pred. 1	512	129	79.88%
Pred. 0	20	89	81.65%
Class recall	96.24%	40.83%	

  Performance on Validation Data:
  Accuracy: **72.80%**

	True 1	*True 0*	*Class Precision*
Pred. 1	154	54	74.04%
Pred. 0	14	28	66.67%
Class recall	91.67%	34.15%	

Among all the above models, based on their performance, the model built with 75% of the training data looks to be more robust than the other models.

After changing the decision tree parameters as below, we were able to build a decision tree (see Appendix D) with an accuracy of 82.13% on the Training data and 75.60% on the Validation data.

- Split criterion: *Gini_Index*
- Maximal Depth: *20*
- Pruning: Yes
- Confidence: *0.5*
- Prepruning: Yes
- Minimal Gain: *0.065*
- Minimal Leaf Size: *5*
- Minimal Size for Split: *5*
- Number of prepruning alternatives: *2*

We could also observe that pruning a decision tree gives a better model in terms of complexity. Without pruning, accuracy of the model tends to be more on the training data, because we allow the tree to expand till it reaches pure nodes/leaves and hence increase the chances of overfitting the data.

While changing the decision tree parameters in building the models with different partitions, below are a few important parameters which seemed to be more influential on the accuracy of the model:

- Minimal Gain: While adjusting the minimal gain value, we observed that the accuracy of the model on the validation data decreases if we either increase its value above 0.66 (unfitting) or decrease its value below 0.64 (overfitting).
- Minimal Leaf Size: Similarly, low leaf size might lead to overfitting the data.
- Split Criterion: For this dataset, except for Gini Index and Information gain, all other criterion doesn't seem to be useful in building a robust tree with a higher depth than 2.

c. Model built using W-SimpleCart:
Performance on Training Data:
Accuracy: **77.47%**

	True 1	*True 0*	*Class Precision*
Pred. 1	485	122	79.90%
Pred. 0	47	96	67.13%
Class recall	91.17%	44.04%	

Performance on Validation Data:
Accuracy: **75.60%**

	True 1	*True 0*	*Class Precision*
Pred. 1	152	45	77.16%
Pred. 0	16	37	69.81%
Class recall	90.48%	45.12%	

- Model built using W-J48:
Performance on Training Data:
Accuracy: **81.20%**

	True 1	*True 0*	*Class Precision*
Pred. 1	467	76	86.00%
Pred. 0	65	142	68.60%
Class recall	87.78%	65.14%	

Performance on Validation Data:
Accuracy: **74.00%**

	True 1	*True 0*	*Class Precision*
Pred. 1	147	44	76.96%
Pred. 0	21	38	64.41%
Class recall	87.50%	46.34%	

After adjusting the decision tree parameters, above are the best accuracy values obtained while using CART and J48 decision trees models. We could also see that among the two, the model built using **CART** had a better accuracy value.

d. Based on the performance and relative adjustment of a few parameters, we have considered the model built using CART (see Appendix E), to be our best model. Further, while analyzing the stability of our model using different training samples, we could observe the below changes in the accuracy levels of our model.

Local Random Seed	Accuracy on Training Data	Accuracy on Validation Data
10	75.6%	70.00%
100	81.07%	77.60%
1000	77.73%	72.40%
10000	80.67%	67.60%
No local Random Seed	77.47%	75.60%

We could observe that the accuracy of these models vary with different seeds and thus add to the reference that decision trees are unstable with slight changes in the training dataset. However, when we observe the trees in each model, we could see that most of the trees look similar at the top.

Answer #4

Based on the costs provided, below are the miscalculation costs incurred by our model:

Threshold	Costs on Training Data	Costs on Validation Data
0.7	87.600 DM	96.400 DM
0.75	63.733 DM	50.000 DM

We could see that when we increase the classification threshold, the miscalculation costs decreases.

Answer #5

a. The tree in the best model that we have decided on has a depth of 8. It contains 23 nodes out of which 12 are leaf nodes. A comparison of the variables at the top (5) of the tree in our best model to that of the model built in Q2 are as follows:

Best Model	*Model in Q2*
Chk_Acc	Chk_Acc
Duration	History
History	Amount
Sav_Acc	Duration
Used_Car	Install_Rate

We could see that there are significant differences in the variables at the top of each tree.

b. From the tree structure in our best model, we could see that below are the two relatively pure leaf nodes.

```
CHK_ACCT=(0)|(1)
|  DURATION < 22.5
|  |  HISTORY=(1)|(0): 0(21.0/7.0)
|  |  HISTORY!=(1)|(0)
|  |  |  AMOUNT < 7491.5
|  |  |  |  DURATION < 11.5: 1(66.0/12.0)
|  |  |  |  DURATION >= 11.5
|  |  |  |  |  AMOUNT < 1387.5
|  |  |  |  |  |  GUARANTOR=(0)
|  |  |  |  |  |  |  NUM_CREDITS < 1.5: 0(32.0/16.0)
|  |  |  |  |  |  |  NUM_CREDITS >= 1.5: 1(10.0/4.0)
|  |  |  |  |  |  GUARANTOR!=(0): 1(10.0/1.0)      ← Relatively Pure leaf node in the tree
|  |  |  |  |  AMOUNT >= 1387.5: 1(90.0/30.0)
|  |  |  AMOUNT >= 7491.5: 0(6.0/1.0)             ← Relatively Pure leaf node in the tree
|  DURATION >= 22.5
|  |  SAV_ACCT=(0)|(1)|(2)
|  |  |  DURATION < 47.5
|  |  |  |  USED_CAR=(0): 0(85.0/52.0)
|  |  |  |  USED_CAR!=(0): 1(17.0/6.0)
|  |  |  DURATION >= 47.5: 0(31.0/5.0)
|  |  SAV_ACCT!=(0)|(1)|(2): 1(29.0/12.0)
CHK_ACCT!=(0)|(1): 1(397.0/60.0)
```

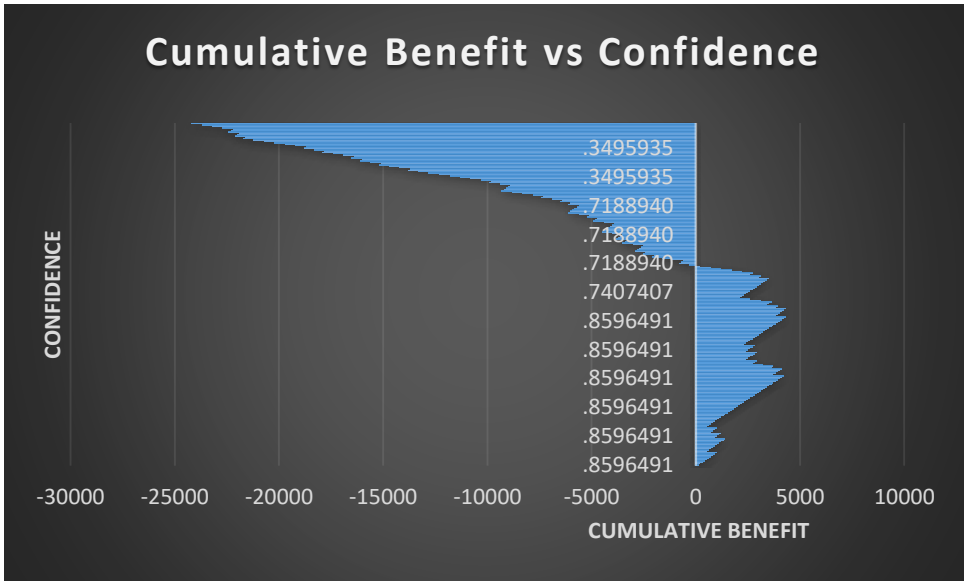
- Below are the probabilities for 'Good' and 'Bad' in the above pure leaf nodes.

Pure Leaf Node	Probability for 'Good'	Probability for 'Bad'
1(10.0/1.0)	0.909	0.09
0(6.0/1.0)	0.143	0.857

c. Two sample rules obtained from the tree are:
   1. IF (Chk_Acct == 0 OR 1) AND (Duration < 22.5) AND (History == 0 OR 1) THEN Class=0
   2. IF (Chk_Acct != 0 OR 1) THEN Class=1

Answer #6

After sorting the predicted probabilities, we observed that the maximum benefit that the model could achieve was 4300 DM at a confidence level of 0.8596. Based on these values, we would recommend a cutoff value of 0.85 while using our model.



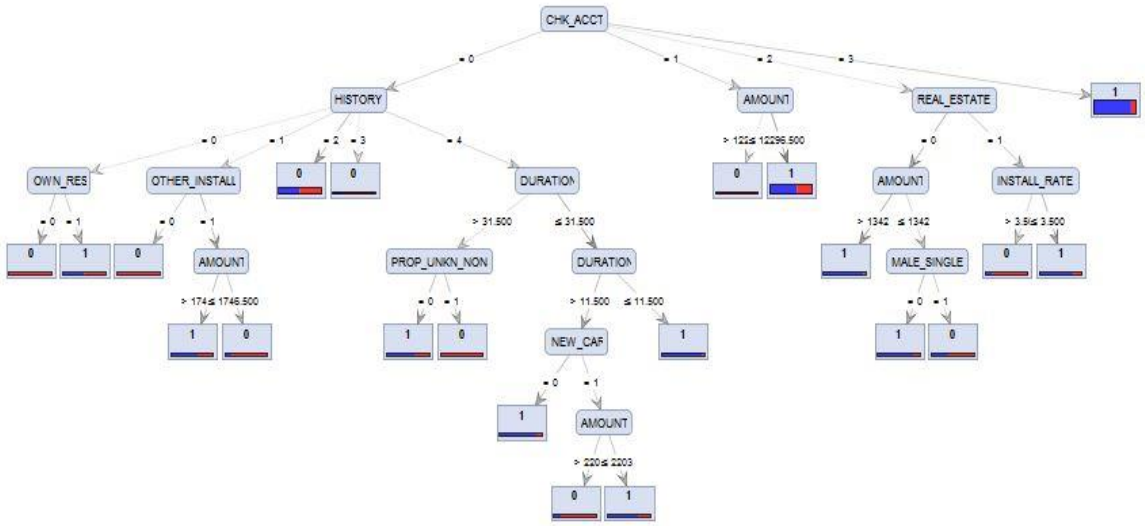
Appendix

Appendix A

Categorical Variable descriptions:

1. Checking Account
    - 0: < 0 DM
    - 1: 0 < … < 200 DM
    - 2: => 200 DM
    - 3: no checking account
2. History
    - 0: no credits taken
    - 1: all credits at this bank paid back duly
    - 2: existing credits paid back duly till now
    - 3: delay in paying off in the past
    - 4: critical account
3. Savings Account
    - 0: < 100 DM
    - 1: 100 <= … < 500 DM
    - 2: 500 <= … < 1000 DM
    - 3: => 1000 DM
    - 4: unknown/ no savings account
4. Employment
    - 0: unemployed
    - 1: < 1 year
    - 2: 1 <= … < 4 years
    - 3: 4<= … < 7 years
    - 4: >= 7 years
5. Present Resident
    - 0: <= 1 year
    - 1: < … <= 2 years
    - 2: < … <= 3 years
    - 3: > 4 years
6. Job
    - 0: Unemployed/Unskilled – Non resident
    - 1: Unskilled - resident
    - 2: Skilled Employee/Official
    - 3: Management/Self-employed/highly qualified employee/officer

Appendix B

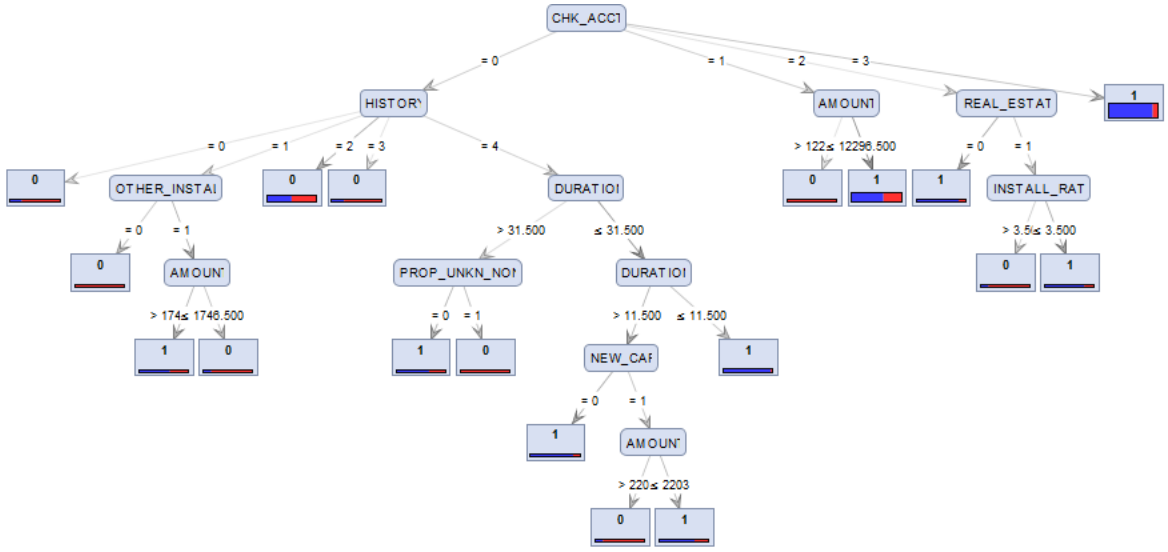


Appendix C

Accuracy: **75.50%**

	True 1	*True 0*	*Class Precision*
Pred. 1	614	159	79.43%
Pred. 0	86	141	62.11%
Class recall	87.71%	47.00%	

Appendix D



Performance on Training Data (75%):

Accuracy: **82.13%**

	True 1	*True 0*	*Class Precision*
Pred. 1	503	105	82.73%
Pred. 0	29	113	79.58%
Class recall	94.55%	51.83%	

Performance on Validation Data (25%):

Accuracy: **75.60%**

	True 1	*True 0*	*Class Precision*
Pred. 1	153	46	76.88%
Pred. 0	15	36	70.59%
Class recall	91.07%	43.90%	