9/19/2015

Aditi Vishwasrao - 674849041

Mounica Sirineni - 652518967

Nishanth Reddy Konkala - 658824939

IDS 572 Assignment 1

**1) Explore the data:**

**What is the proportion of “Good” to “Bad” cases?**

The data has 1000 observations. Out of these 1000, 700 are “Good” cases and 300 are “Bad” cases i.e., 700 observations have response as “1” and 300 have response as “0”.

Therefore, the *proportion of “Good” to “Bad” cases is 700/300 =* ***2.33***.

**Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-values attributes, frequencies of different category values.**

Data has 32 variables. Excluding Obs# and Response, it has 30 independent variables. Out of these 30, 6 variables are real valued attributes and 24 are categorical attributes.

**The following table provides the descriptions of real valued attributes.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Missing** | **Minimum** | **Maximum** | **Average** | **Deviation** |
| **DURATION** | 0 | 4 | 72 | 20.903 | 12.059 |
| **AMOUNT** | 0 | 250 | 18424 | 3271.258 | 2822.737 |
| **INSTALL\_RATE** | 0 | 1 | 4 | 2.973 | 1.119 |
| **AGE** | 0 | 19 | 75 | 35.546 | 11.375 |
| **NUM\_CREDITS** | 0 | 1 | 4 | 1.407 | 0.578 |
| **NUM\_DEPENDENTS** | 0 | 1 | 2 | 1.155 | 0.362 |

**The following table provides the frequency of categorical attributes.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Type** | **Missing** | **Least (frequency)** | **Most (frequency)** | **Frequency** |
| **CHK\_ACCT** | Nominal | 0 | 2 (63) | 3 (394) | 3-394  0-274  1-269  2-63 |
| **HISTORY** | Nominal | 0 | 0 (40) | 2 (530) | 2-530  4-293  3-88  1-49  0-40 |
| **NEW\_CAR** | Binominal | 0 | 1 (234) | 0 (766) | 0-766  1-234 |
| **USED\_CAR** | Binominal | 0 | 1 (103) | 0 (897) | 0-897  1-103 |
| **FURNITURE** | Binominal | 0 | 1 (181) | 0 (819) | 0-819  1-181 |
| **RADIO/TV** | Binominal | 0 | 1 (280) | 0 (720) | 0-720  1-280 |
| **EDUCATION** | Binominal | 0 | 1 (50) | 0 (950) | 0 - 950  1 - 50 |
| **RETRAINING** | Binominal | 0 | 1 (97) | 0 (903) | 0 - 903  1 - 97 |
| **SAV\_ACCT** | Nominal | 0 | 3 (48) | 0 (603) | 0 - 603  4 - 183  1 - 103  2 - 63  3 – 48 |
| **EMPLOYMENT** | Nominal | 0 | 0 (62) | 2 (339) | 2 - 339  4 - 253  3 - 174  1 - 172  0 – 62 |
| **MALE\_DIV** | Binominal | 0 | 1 (50) | 0 (950) | 0 – 950  1 - 50 |
| **MALE\_SINGLE** | Binominal | 0 | 0 (452) | 1 (548) | 1 – 548  0 – 452 |
| **MALE\_MAR\_or\_WID** | Binominal | 0 | 1 (92) | 0 (908) | 1 – 908  0 – 92 |
| **CO-APPLICANT** | Binominal | 0 | 1 (41) | 0 (959) | 0 – 959  1 – 41 |
| **GUARANTOR** | Binominal | 0 | 1 (52) | 0 (948) | 0 – 948  1 – 52 |
| **PRESENT\_RESIDENT** | Nominal | 0 | 1 (130) | 4 (413) | 4 - 413  2 - 308  3 - 149  1 – 130 |
| **REAL\_ESTATE** | Binominal | 0 | 1 (282) | 0 (718) | 0 – 718  1 – 282 |
| **PROP\_UNKN\_NONE** | Binominal | 0 | 1 (154) | 0 (846) | 0 – 846  1 – 154 |
| **OTHER\_INSTALL** | Binominal | 0 | 1 (186) | 0 (814) | 0 – 814  1 – 186 |
| **RENT** | Binominal | 0 | 1 (179) | 0 (821) | 0 – 821  1 – 179 |
| **OWN\_RES** | Binominal | 0 | 0 (287) | 1 (713) | 1 – 713  0 – 287 |
| **JOB** | Nominal | 0 | 0 (22) | 2 (630) | 2 - 630  1 - 200  3 - 148  0 – 22 |
| **TELEPHONE** | Binominal | 0 | 1 (404) | 0 (596) | 0 – 596  1 – 404 |
| **FOREIGN** | Binominal | 0 | 1 (37) | 0 (963) | 0 – 963  1 - 37 |

**Anything noteworthy in the data? Which variables do you think will be most relevant for the outcome of interest? (Why?)**

The following is noteworthy in the data:

* There are no missing values

The following is the analysis of various variables in the data set. Clicking on the hyperlink will display the respective graphs at the end of this document.

* [*Checking Account status*](#Checking_Account)*:*Maximum number of applicants (73%) have the checking amounts between 1000DM and 5000 DM. 54% of these 73% applicants have good credit rating. Most of the respondents (34.50%) who had a good credit rating do not have a checking account.
* [*Employment:*](#Job)Approx 63% of the applicants are skilled professionals. 44.40% of them have good credit rating. Only about 2% of the applicants are unemployed.
* [*Duration:*](#Duration)Duration of credit is right skewed distribution with majority of observations in the range of 4-28 months.
* [*Used Car:*](#Used_Car)Only 8.6% of the respondents have a used car and most of them have a good credit rating. About 61.40% applicants do not have used car and have good credit rating.
* [*Credit History:*](#History)Most respondents have paid back the credit duly till now. However among these there is a high proportion of people who have a bad credit rating. The proportion of good credit rating is higher in the respondents who have a critical account.
* [*Education:*](#Education)Most of the respondents have claimed that they are uneducated. The proportion of good credit rating is however quite good among the uneducated respondents.
* [*Savings Account:*](#Saving_Account)Majority of applicants have < 100 DM in their savings account. The proportion of good to bad credit rating is more among those who have > 1000 DM in their saving account.
* [*Foreigners:*](#Foreign)66.70% of Resident applicants have good credit rating.
* [*Other Install:*](#Other_Installment)81.40% of applicants do not have other installment plan credit and 59% of them have good credit rating.

The following variables show a weak relationship with the credit rating of an applicant. We have chosen to display only one graph for the variable [*telephone*](#Telephone)in order to demonstrate the trend in these variables. Their distribution in the dataset is as follows:

***New Car:***About 77% of the observations do not have a new car.

***Furniture:*** About 82% of the observations do not have purpose of credit as Furniture.

***Radio/TV:*** About 72% of the observations do not have purpose of credit as Radio/TV.

***Retraining:*** About 6% of the observations have purpose of credit as Retraining.

***Male\_Div:*** Only 5% of applicants for credit are divorced males.

***Male\_Mar\_or\_Wid:*** About 9% of applicants are married or widowed males.

**Variables of interest:**

Based on the above graphs and description, the following variables are of interest.

* Chk\_Acct
* History
* Amount
* Duration
* Other\_Install

These variables have a strong relationship with the credit rating of applicants.

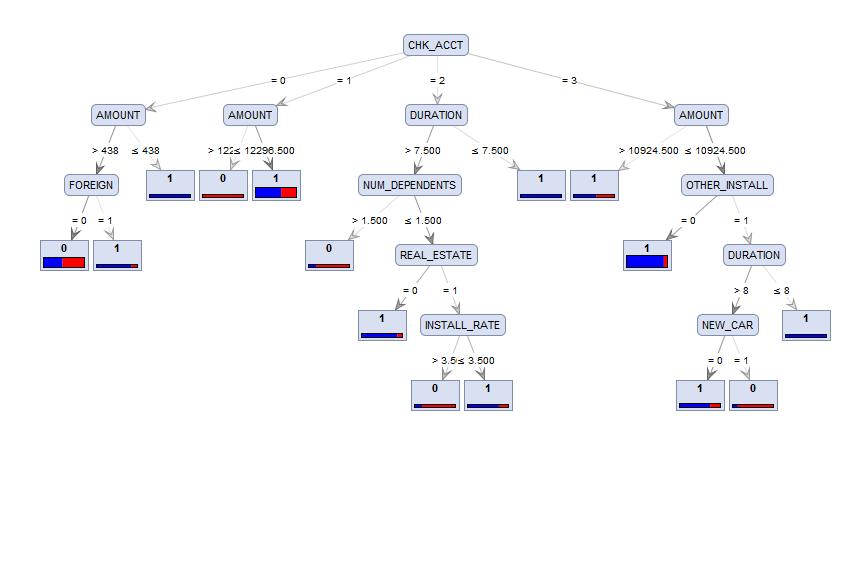
**Question 2:**

**We will first focus on a descriptive model – i.e. assume we are not interested in prediction. Develop a decision tree on the full data.**

**Answer:**

The RapidMiner process of developing a decision tree on full data is [illustrated here](#Full_data_decision)

The screenshot for the parameter considered for this tree is [shown here](#Parameters_Full_data_decision)



**Which variables are used to differentiate the good cases from the bad ones?**

|  |  |
| --- | --- |
| **Split Criteria** | **Variables used to differentiate ‘good’ from ‘bad’ cases** |
| Information Gain | CHK\_ACCT, HISTORY, AMOUNT, OTHER\_INSTALL |
| Gini Index | CHK\_ACCT, HISTORY, OWN\_RES, DURATION |
| Gain Ratio | AMOUNT, AGE, DURATION, INSTALL RATE |

**What levels of accuracy/error are obtained? What is the accuracy for the “good” and “bad” cases?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **Accuracy of Bad Cases** | **Accuracy of good cases** |
| **Information Gain** | 94 | 94.12 | 85.33 | 94.12 | 93.96 |
| **Gini Index** | 93.6 | 93.38 | 84.67 | 93.38 | 93.68 |
| **Gain Ratio** | 80.60 | 71.20 | 59.33 | 71.20 | 83.73 |

**Do you think this is a reliable (robust?) description?**

A robust model should give nearly same accuracy with the change in the parameters. But we observed that with ‘Information Gain’ as splitting criteria and by changing the ‘confidence’ and ‘minimal gain’ we see that the accuracy is changing significantly. Here, the data is not split into training and validation data. We can determine the reliability of the model only when the model is divided into Training and Validation data and the training data is tested upon the validation data.

**What decision tree node parameters do you use to get a good model (and why?)**

**Split Criteria:** Information gain since the accuracy is 94%.

**Pruning:** Pruning was applied in order to avoid over fitting of data and to make the model less complex on the validation data. If pruning was not applied, the size of the tree increases and decreases accuracy.

**Confidence:** Default value (0.25) of confidence was used. If confidence was decreased, there was an increase in accuracy, but the goal of developing this model is to perform better on unseen data. Therefore confidence value was not reduced.

**Minimum size for split:** We didn’t want the model to split further if the number of observations in a node is less than 2 so as to avoid complexity.

**Minimal Leaf Size:** This is set to 2, to split the nodes if the number of observations are >=2.

**No. Of Pre pruning alternatives:** Changing this didn’t show any effect.

**Question 3:**

**We next consider developing a model for prediction. For this, we should divide the data into Training and Validation sets.**

The screenshot for [Validation process](#Validation_process) and the [Validation model](#Training_and_vaidation_set) are included in the appendix.

**Model Performance**

**Parameters:**[(screenshot)](#param_Performance_eval_model)

Maximal Depth: 20

Apply Pruning: Checked

Confidence: 0.25

Apply Prepruning: Checked

Minimal gain: 0.05

Minimal leaf size: 4

Minimal size for split: 4

Number of prepruning alternatives: 3

**a. Consider a partition of the data into 50% for Training and 50% for Test. What model performance do you obtain? Is the model reliable (why or why not)?**

|  |  |  |  |
| --- | --- | --- | --- |
| **50-50 Split Ratio** |  | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| **Information Gain** | No Random Seed | 81.20 | 68.80 |
| Random Seed-400 | 87.00 | 69.00 |
| Random Seed-500 | 87.20 | 71.60 |
| Random Seed-600 | 87.40 | 70.80 |
| **Gini Index** | No Random Seed | 81.00 | 69.00 |
| Random Seed-400 | 84.80 | 68.40 |
| Random Seed-500 | 85.60 | 69.80 |
| Random Seed-600 | 88.40 | 70.60 |
| **Gain Ratio** | No Random Seed | 72.60 | 69.00 |
| Random Seed-400 | 70.20 | 71.60 |
| Random Seed-500 | **76.80** | **71.20** |
| Random Seed-600 | 74.20 | 68.00 |

In the second question, we have seen that Information gain based split criteria has given us higher accuracy when compared to the others and the same trend followed in the training data set. But a reliable and robust model must have training and validation accuracies with small differences. From the above table we can see that, ‘Information Gain’ and ‘Gini Index’ based splits are giving training and validation accuracies with large differences (>10%) and some splits in these have validation accuracies less than 70%. (A model with >=70% in both training and validation is generally considered to be reliable).

The model highlighted in the above table with split criteria as gain ratio and random seed as 500 is reliable because:

* There is even split in training and validation data.
* The performance in training and validation doesn’t differ hugely (76.8-71.2=5.6).
* Training performance > validation performance and changes in the parameters didn’t give much different values.

As seen in question 2, we increased the ‘minimal gain’, ‘minimal leaf size’, and ‘minimal size for split’ values in order to make the model more accurate and less complex.

**b. Consider partitions of the data into 70% for Training and 30% for Test, and 80% for Training and 20% for Test and report on model and performance comparisons. Feel free to experiment with other size partitions on the data. Is there any specific model you would prefer for implementation?**

**Parameters:**[(screenshot)](#param_Performance_eval_model)

Maximal Depth: 20

Apply Pruning: Checked

Confidence: 0.25

Apply Prepruning: Checked

Minimal gain: 0.05

Minimal leaf size: 4

Minimal size for split: 4

Number of prepruning alternatives: 3

|  |  |  |  |
| --- | --- | --- | --- |
| **70-30 Split Ratio** |  | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| **Information Gain** | No Random Seed | 84.43 | 68.00 |
| Random Seed-400 | 80.43 | 72.33 |
| Random Seed-500 | 85.14 | 72.00 |
| Random Seed-600 | 81.86 | 71.00 |
| **Gini Index** | No Random Seed | 85.43 | 69.33 |
| Random Seed-400 | 83.29 | 74.33 |
| Random Seed-500 | 86,57 | 72.33 |
| Random Seed-600 | 87.00 | 68.00 |
| **Gain Ratio** | No Random Seed | 72.43 | 64.33 |
| Random Seed-400 | 71.86 | 70.00 |
| Random Seed-500 | 77.00 | 69.00 |
| Random Seed-600 | 72.57 | 72.33 |

|  |  |  |  |
| --- | --- | --- | --- |
| **80-20 Split Ratio** |  | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| **Information Gain** | No Random Seed | 85.38 | 72.50 |
| Random Seed-400 | 82.12 | 68.50 |
| Random Seed-500 | 78.50 | 67.50 |
| Random Seed-600 | 82.50 | 71.50 |
| **Gini Index** | No Random Seed | 85.00 | 72.00 |
| Random Seed-400 | 82.62 | 72.00 |
| Random Seed-500 | 83.50 | 69.00 |
| Random Seed-600 | 83.25 | 71.50 |
| **Gain Ratio** | No Random Seed | 70.88 | 66.50 |
| Random Seed-400 | 72.38 | 69.00 |
| Random Seed-500 | 73.75 | 66.50 |
| Random Seed-600 | 73.00 | 70.50 |

In both the 70-30 and 80-20 splits, we observe that the training and validation performances are above 70 with minimum differences for ‘Gain Ratio’ and ‘random seed’ as 600. These models seem to be reliable as there are no huge differences with iterations (change in random seed), but we do not prefer to implement any of these because of the split ratio. If the training data ratio is high, the model will over fit the training data and will become complex for the validation data.

**In developing the models above, change some of the decision tree options and see if and how they affect performance (for example, the minimum number of cases at a leaf node, the split criteria).**

**Parameters:**[(screenshot)](#Parameters_Full_data_decision)

Maximal Depth: 20

Apply Pruning: Checked

Confidence: 0.5

Apply Prepruning: Checked

Minimal gain: 0.01

Minimal leaf size: 2

Minimal size for split: 2

Number of prepruning alternatives: 3

|  |  |  |  |
| --- | --- | --- | --- |
| **70-30 Split Ratio** |  | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| **Information Gain** | No Random Seed | 93.57 | 65.67 |
| Random Seed-400 | 93.29 | 70.67 |
| Random Seed-500 | 92.29 | 72.00 |
| Random Seed-600 | 92.43 | 69.00 |
| **Gini Index** | No Random Seed | 94.00 | 64.33 |
| Random Seed-400 | 92.00 | 70.67 |
| Random Seed-500 | 92.14 | 70.67 |
| Random Seed-600 | 92.86 | 67.33 |
| **Gain Ratio** | No Random Seed | 88.29 | 69.33 |
| Random Seed-400 | 89.29 | 67.67 |
| Random Seed-500 | 80.57 | 67.00 |
| Random Seed-600 | 82.57 | 73.00 |

|  |  |  |  |
| --- | --- | --- | --- |
| **80-20 Split Ratio** |  | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| **Information Gain** | No Random Seed | 93.62 | 69.00 |
| Random Seed-400 | 93.88 | 71.50 |
| Random Seed-500 | 94.50 | 67.50 |
| Random Seed-600 | 95.00 | 65.00 |
| **Gini Index** | No Random Seed | 93.38 | 68.50 |
| Random Seed-400 | 92.38 | 70.00 |
| Random Seed-500 | 93.62 | 68.50 |
| Random Seed-600 | 94.12 | 69.50 |
| **Gain Ratio** | No Random Seed | 88.75 | 70.50 |
| Random Seed-400 | 82.50 | 72.00 |
| Random Seed-500 | 87.75 | 69.50 |
| Random Seed-600 | 85.50 | 68.50 |

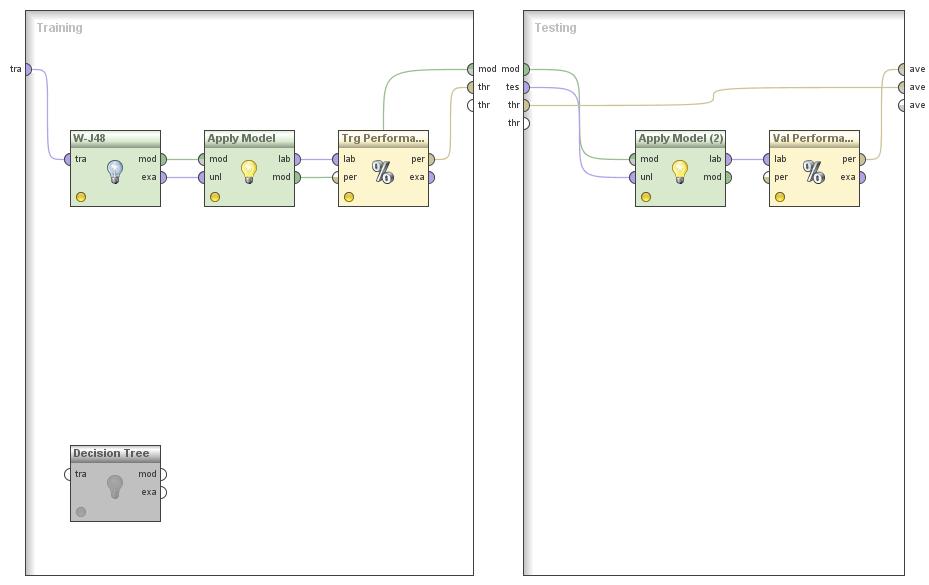
By increasing confidence and decreasing the ‘minimal gain’, ‘minimal leaf size’ we observe that there is increase in the accuracy of the training performance in all the split criteria’s but there is no significant increase in the validation performances.

**Also, does pruning give a better model – please explain why or why not? Which parameter values do you find to be useful – are they the same for different training-test partitions?**

Pruning does give us better model because it decreases the level of complexity in for the validation data set. It also decreases the depth of the tree, which also helps in getting a better model. We found that the following [parameters](#param_Performance_eval_model) were useful in getting a good model and they are same across different split criteria’s except for the random seed.

**c. Also, consider two other type of decision tree operators – for example, CART, J48 – play around with the parameters till you get a ‘good’ model. Do you see any performance differences across different types of decision tree learners?**

**J48**

****

**Parameters:**

Unpruned tree (U): Unchecked

Confidence (C): 0.25

Minimum number of instances per leaf (M): 2

Reduced error pruning (R): Unchecked

Number of folds for reduced error pruning (N): Blank

Binary splits (B): Unchecked

Subtree raising (S): Unchecked

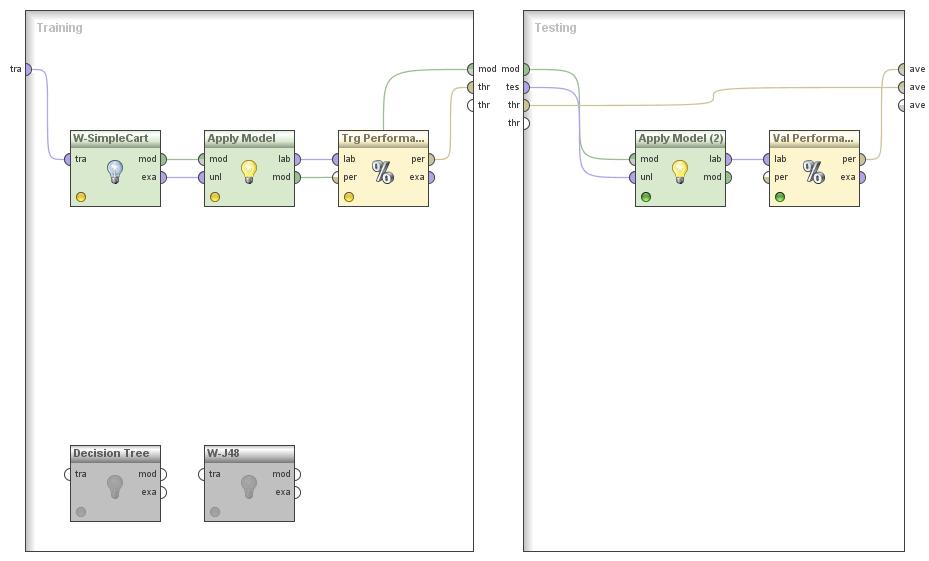
Do not clean up after tree has been built (L): Unchecked

Laplace smoothing for predicted probabilities (A): Unchecked

Seed for random data shuffling (Q): Blank

|  |  |  |
| --- | --- | --- |
| **50-50 Split Ratio** | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| No Random Seed | 88.40 | 71.00 |
| Random Seed-400 | 84.40 | 72.60 |
| Random Seed-500 | 87.00 | 70.20 |
| Random Seed-600 | 91.40 | 67.60 |

**CART:**

****

**Parameters:**

Random number seed (S): 1.0

Debug mode (D): Unchecked

Minimal number of instances at the terminal nodes (M): 2.0

Number of folds used in the minimal cost-complexity pruning (N): 5.0

Don’t use the minimal cost-complexity pruning (U): Unchecked

Don’t use the heuristic method for binary split (H): Unchecked

Use 1 SE rule to make pruning decision (A): Unchecked

Training data size (C): 1.0

|  |  |  |
| --- | --- | --- |
| **50-50 Split Ratio** | **Training Accuracy (%)** | **Validation Accuracy (%)** |
| No Random Seed | 79.00 | 74.00 |
| Random Seed-400 | 79.60 | 70.40 |
| Random Seed-500 | 69.60 | 70.40 |
| Random Seed-600 | 83.20 | 71.60 |

The model’s behavior does change with different decision trees. But the difference between the accuracies in the training and validation data sets is not very different.

It looks like the CART model is more reliable than J48 because the difference between the training and validation accuracies in each case is lesser in the CART model.

**d. Decision tree models are referred to as ‘unstable’ – in the sense that small differences in training data can give very different models. After selecting a set of parameters which you find to work well, try building different models with different training samples (you can change the random seed for this). Do you find your models to be unstable? Are there similarities in, say, the upper part of the tree – and what does this indicate?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Gain Ratio** | No Random Seed | 72.60 | 69.00 |
| Random Seed-400 | 70.20 | 71.60 |
| Random Seed-500 | **76.80** | **71.20** |
| Random Seed-600 | 74.20 | 68.00 |

For our best model for 50-50 split ratio, in the above table you see that there are differences in the performance with different training data. But that difference is not very huge and significant. With different values of random seed, there is no difference in the upper part of the tree. We had CHK\_ACT, AMOUNT, HISTORY and duration for all the random seed values. This indicates that the change in the training data affects only the lower part of the tree when leaf nodes arrive.

**4. Use the misclassification costs to assess performance of a chosen model from 3 above. Examine how different cutoff values for classification threshold make a difference – what do you find?**

The misclassification cost for various thresholds is as follows:

|  |  |  |
| --- | --- | --- |
| **Threshold** | **Training Cost Performance** | **Validation Cost Performance** |
| 0.0 | 144 | 143 |
| 0.1 | 144 | 143 |
| 0.2 | 139.2 | 143 |
| 0.3 | 139.2 | 143 |
| 0.4 | 139.2 | 94.4 |
| 0.5 | 82.6 | 94.4 |
| 0.6 | 81.5 | 94.4 |
| 0.7 | 51.4 | 67.4 |
| 0.8 | 50.4 | 67.4 |
| 0.9 | 50.7 | 74.6 |
| 1.0 | 70 | 70.4 |

The misclassification cost decreases as the threshold is increased. However, at the extreme value of 1, the cost performance shows a rising trend again. This is illustrated by the following line graph.

We can see from the graph that the values nearer to the middle threshold value of 0.5 have



**5. Let’s examine your ‘best’ decision tree model obtained.**

**(a) What is the tree depth? And how many nodes does it have? What are the variables towards the ‘top’ of the tree, and are they similar to what you found in Question 2?**

Tree depth: **6**

Number of nodes: **23**

Variables towards the top of the tree: **CHK\_ACCT, AMOUNT, DURATION.** These are same as the ones we found in Question 2.

**(b) Identify two relatively pure leaf nodes. What are the ‘probabilities for ‘Good’ and ‘Bad’ in these nodes?**

Relatively pure leaf nodes:

Case 1:

CHK\_ACCT = 3

| AMOUNT > 10924.500: 1 {1=4, 0=3}

| AMOUNT ≤ 10924.500

| | OTHER\_INSTALL = 0: 1 {1=300, 0=25}

Probability for Good: **0.923**

Probability for Bad : **0.077**

Case 2:

CHK\_ACCT = 1

| AMOUNT > 12296.500: 0 {1=0, 0=12}

Probability for Good: **0**

Probability for Bad : **1**

**(c) The tree can be used to obtain rules – give two sample rules obtained from the tree. (Rules will be of the form IF condition AND condition AND…. THEN classification).**

IF CHK\_ACCT = 0 AND AMOUNT ≤ 438

RESPONSE = 1

IF CHK\_ACCT = 3 AND AMOUNT ≤ 10924.500 AND OTHER\_INSTALL = 0

THEN RESPONSE = 1

**6.**

**How far into the validation data would you go to get maximum net benefit? In using this model to score future credit applicants, what cutoff value for predicted probability would you recommend? Provide appropriate performance values to back up your recommendation.**

When sorted as per decreasing probabilities, the 214th observation gives the maximum net benefit of 4500. The recommended cut off value for predicted probability would be 0.8571. For the Cost performance plotted in the graph below, we can see that the peak of the net benefits value is at a predictive probability of **0.8571.**



The following Excel sheet shows the calculation of the net cumulative benefits and the profit and loss.



**APPENDIX**

**Amount Vs Response**



We can see that maximum number of applicants (73%) have the credit amounts between 1000DM and 5000 DM. 54% of them have good credit rating.

**Job Vs Response**



A majority of the applicants (approx 63%) are skilled professionals. 44.40% of them have good credit rating. Only about 2% of the applicants are unemployed.

**Used Car Vs Response**



Only 8.6% of the respondents have a used car and most of them have a good credit rating. About 61.40% applicants do not have used car and have good credit rating.

**Checking Account Vs Response**



Most of the respondents (34.50%) who had a good credit rating do not have a checking account.

**History Vs Response**



Most respondents have paid back the credit duly till now. However among these there is a high proportion of people who have a bad credit rating. The proportion of good credit rating is higher in the respondents who have a critical account.

**Education Vs Response**



Most of the respondents have claimed that they are uneducated. The proportion of good credit rating is however quite good among the uneducated respondents.

**Saving Account Vs Response:**



The proportion of good to bad credit Majority of applicants have < 100 DM in their savings account. rating is more among those who have > 1000 DM in their saving account.

**Residential status Vs Response:**



66.70% of Resident applicants have good credit rating

**Other Installment plans Vs Response:**



81.40% of applicants do not have other installment plan credit and 59% of them have good credit rating.

**Telephone ownership Vs Response:**



**Duration Vs Response**



38% of the respondents had less than 10 months of duration in credit.

**PerformanceVector (Gain Ratio)**

PerformanceVector:

accuracy: 80.60%

ConfusionMatrix:

True: 1 0

1: 628 122

0: 72 178

precision: 71.20% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 628 122

0: 72 178

recall: 59.33% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 628 122

0: 72 178

AUC (optimistic): 0.915 (positive class: 0)

AUC: 0.852 (positive class: 0)

AUC (pessimistic): 0.789 (positive class: 0)

# PerformanceVector (Information Gain)

PerformanceVector:

accuracy: 94.00%

ConfusionMatrix:

True: 1 0

1: 684 44

0: 16 256

precision: 94.12% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 684 44

0: 16 256

recall: 85.33% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 684 44

0: 16 256

AUC (optimistic): 0.985 (positive class: 0)

AUC: 0.983 (positive class: 0)

AUC (pessimistic): 0.980 (positive class: 0)

# PerformanceVector (Gini Index)

PerformanceVector:

accuracy: 93.60%

ConfusionMatrix:

True: 1 0

1: 682 46

0: 18 254

precision: 93.38% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 682 46

0: 18 254

recall: 84.67% (positive class: 0)

ConfusionMatrix:

True: 1 0

1: 682 46

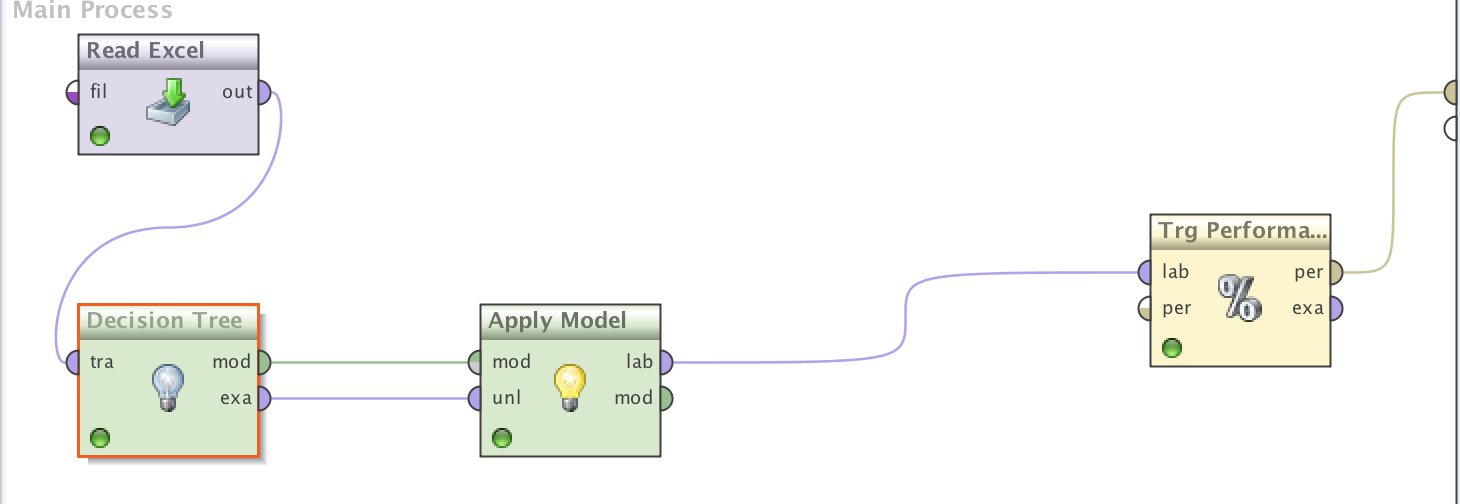
0: 18 254

AUC (optimistic): 0.983 (positive class: 0)

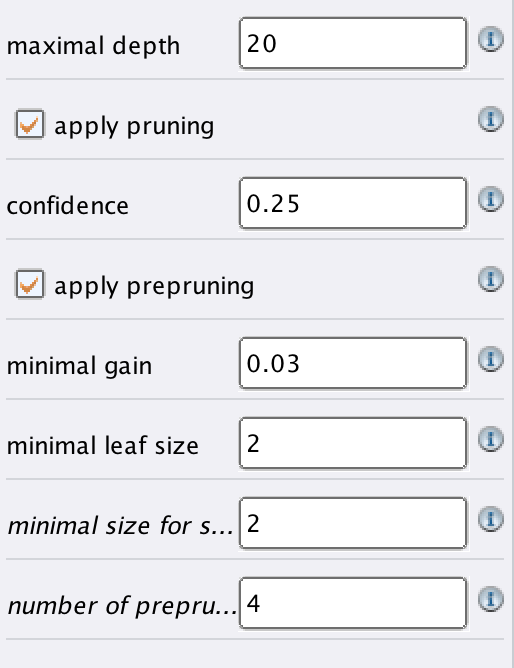
AUC: 0.980 (positive class: 0)

AUC (pessimistic): 0.977 (positive class: 0)

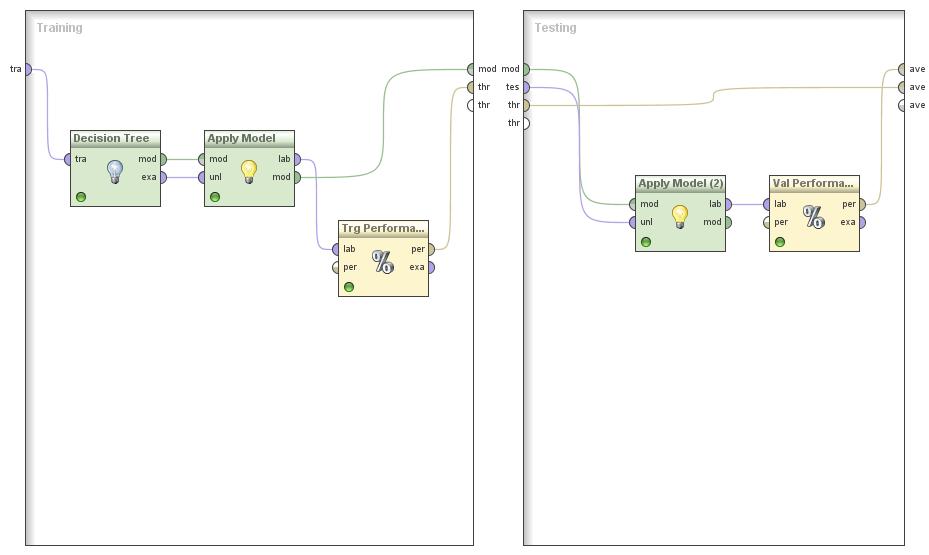
**Developing a decision tree on the full data:**



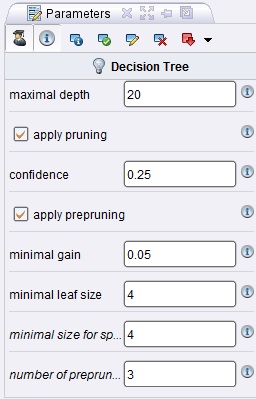
**Parameters for full decision tree:**



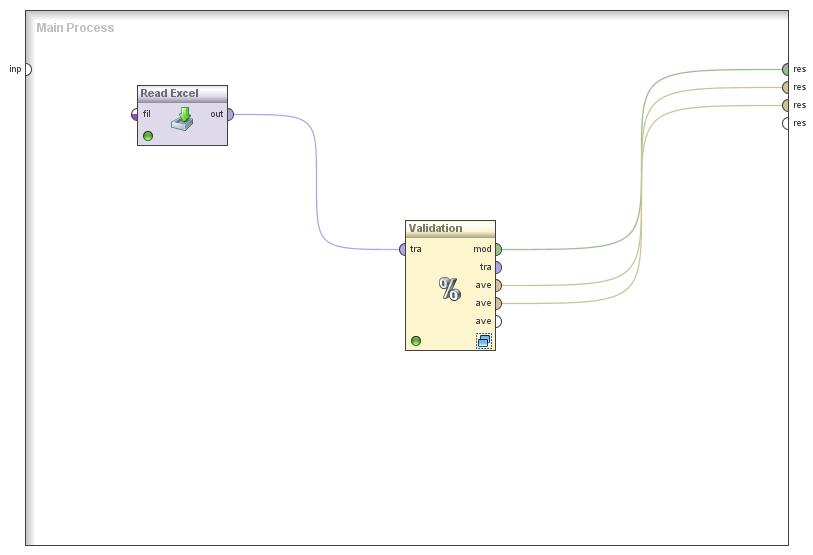
**Model for Performance Evaluation:**



**Parameters for Predictive Models:**



**Process for Validation:**

****

**Training and Validation Set model for prediction**

