Decision Trees

MIRI / MVA - Practice #7 Report

Authors: Cedric Bhihe <cedric.bhihe@gmail.com>
Santi Calvo <s.calvo93@gmail.com>

Delivery: before 2018.05.27 – 23:55

The audit data-set is a simplified financial audit data-set for modelling productive and non-productive audits of a person's financial statement. The dataset is used to illustrate binary classification. A productive audit is one which identifies errors or inaccuracies in the information provided by a client. A non-productive audit is usually an audit which found all supplied information to be in order. The target variable is identified as Adjusted.

1. Import the Audit.xlsx file and convert it to the csv format

Using the R package `xlsx`, we import the xlsx formatted file, convert it to csv format using column ID as row names, before saving it to disk as a csv file. Subsequent calls to the data set, are made by directly loading the csv file in memory.

The data-set contains 2000 observations, and 11 variables beside the boolean target//dependent/response variable "`Adjusted`". 1859 observations have no missing value. 141 have missings distributed according to Figure 1.

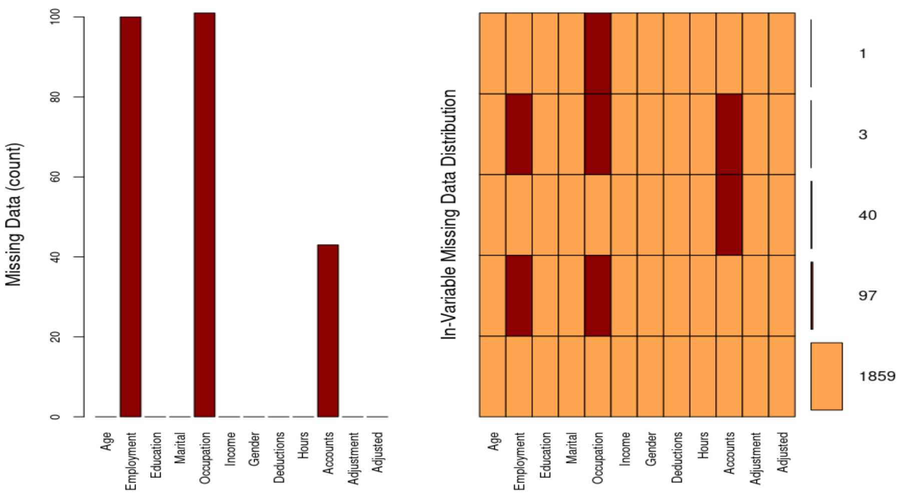

***Fig. 1**: Missings distribution*

We do not perfom imputations as the R package rpart() uses surrogate variables efficiently to handle missings.

2. Decide which predictors you will use. Pre-process the corresponding variables as needed.

We define 10 predictors and 2 supplementary variables as follows:

- 6 active categorical variables: c("Employment","Education", "Marital", "Occupation", "Gender", "Accounts")
- 4 active continuous variables: c("Age","Income","Deductions","Hours")
- 1 supplementary continuous variable: c("Adjustment"), which illustrates a productive audit and `Adjusted=1`.
- 1 supplementary categorical (yes/no or 0/1) variable: c("Adjusted"), which is our target/dependent variable

We pre-process the continuous variables "Age", "Income", "Hours" to transform them in as many categorical variables. Care is taken so each variable's resulting bins (modalities) exhibit similar counts.

"Deductions" remain a continuous variable as its discretization would yield very lopsided bins' frequencies, between two basic modalities: "With" (4.15%) and "Without" (95.85%). See Figure 2.

Table 1 (to the right) summarizes the discretization.

Age	]-inf,27]	]27,38]	]38,50]	]50,inf[
modality	Prime	Middle	Mature	Senior
bin counts	498	558	537	407
Income	]-inf,34.5k]	]34.5k,60k]	]60k,115k]	]115k,inf[
modality	Low	Medium	High	Obscene
bin counts	501	502	503	494

Hours	]-inf,30]	]30,40]	]40,inf[
modality	Part	Reduced	Full
bin counts	344	1063	593

***Table 1**: Discretization of active continuous variables "Age", "Income", "Hours"*

3. Select 1/3 of the data at the end of the data set, as test data.
The training data-set consists of 1333 observations, while the test data-set consists of 667 observations.

4. Build the decision tree to predict var "Adjusted" using training data. Determine cutoff value for optimal decision making.
The fully grown tree for the training dataset using 10 Cross-Validation (CV) replicas, and a complexity parameter of 10^{-3} is shown on Figure 3.

Figure 4 represents the CV normalized error mean (tree cost R(T) and the whole data set based training error as a function of tree size (or α value). The complexity parameter table used to build Figure 4 is available in Appendix B.

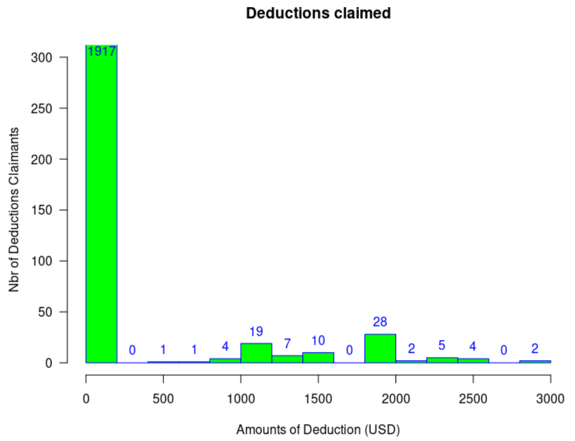

***Fig. 2**: Histogram of deductions claimed (Numbers in blue are counts for each bucket.)*

The red horizontal dashed line represents the minimum tree impurity (MTI) level, and the red dotted line above it MTI + 1 standard error, calculated over the CV errors for value of the complexity parameter, α, in the closed interval [0.001 , 0.150]. The optimum value of α is obtained by inspection. It corresponds to the first value of CV tree cost smaller than MTI + 1 standard error, when scanning normalized mean values of tree impurity starting at root (tree size = 1 for α = 0.150): $\alpha_{opt} = 0.01250$.

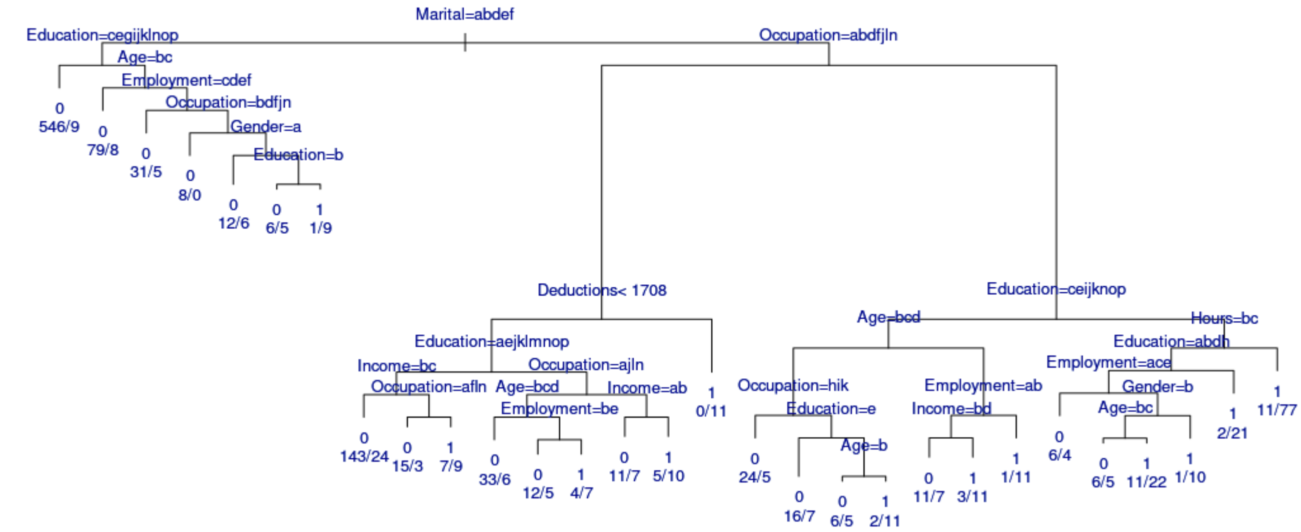

***Fig. 3**: Fully grown decision tree for "Audit" training data-set using 10 Cross-Validation iterations and complexity parameter (cp), $\alpha_{min}=10^{-3}$.*

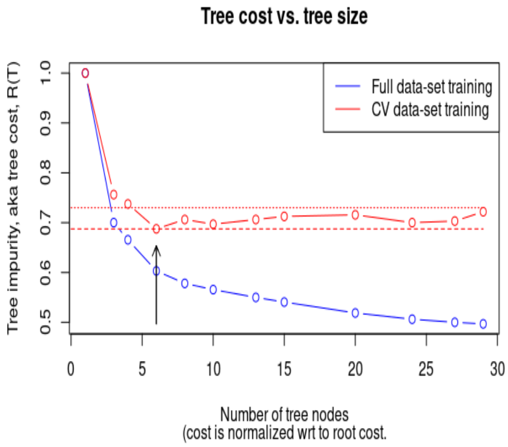

***Fig.4**: Data-sets' training error without (blue) and with (red) cross validation. The red dashed line represents the smallest value of CV training error while the red dotted line above it represents the same + 1 standard error. The black arrow point to the optimum number of nodes for post-pruning.*

This optimum complexity parameter value allows us to post-prune the decision tree at the 5th node split. The result is shown in Figure 5 below along with corresponding split rules.

We notice only one small pure node (for a productive audit, i.e. `Adjusted` =1) in the optimal decision tree, at split 5. The corresponding rules obtained at training are self-explanatory and are in Appendix C.

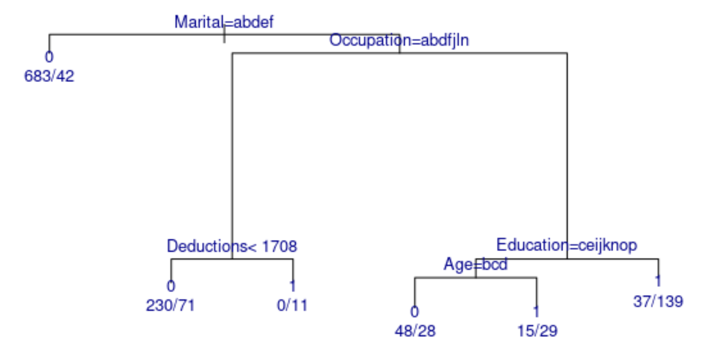

Fig. 5*: Optimally pruned decision tree, with α = 0.01250. Obtained tree size is 6 for 5 splits over 1333 observations.*

5. Plot the importance of variables in the prediction.

Figure 6 shows variables' importance. Most important are `Marital`, `Income` and `Occupation`.

A second group of lesser importance is made of:

- `Education`, `Gender`, and `Age`.

The least important variables are:

- `Hours`, `Deductions`, `Employment`, and `Accounts`.

The variable `Accounts` is the least important. As it is not a predictor of variable `Adjusted`, it would be safe to remove it from the decision-tree analysis, with no effect on the analysis' outcome.

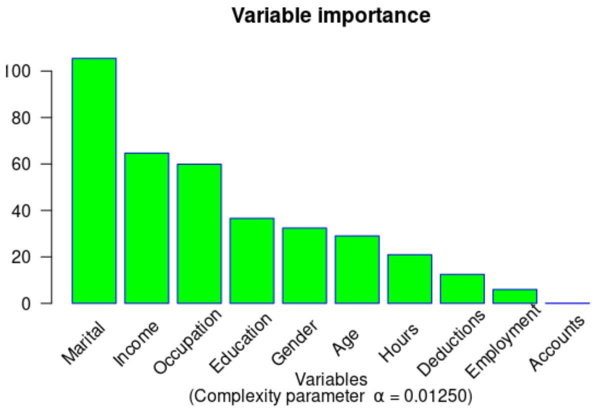

Fig. 6*: Plot of variable importance for an optimally pruned decision tree (α = 0.01250).*

6. Compute the accuracy, precision, recall and AUC on the test data.

7. Perform a Random Forest on test data

Appendix A: Data-set's variables' dictionary

ID	Unique identifier for the person's being audited.
Age	Age of the person being audited.
Employment	Type of employment.
Education	Highest level of education.
Marital	Current marital status.
Occupation	Type of occupation.
Income	Amount of income declared.
Gender	Person's gender.
Deductions	Total amount of expenses that a person claims in their financial statement.
Hours	Average number of hours worked per week.
Accounts	Country in which the person has most of their money banked.
Adjustment	Monetary amount of any adjustment to the person's financial claims as a result of a productive audit. This variable is thus a measure of the size of the risk associated with the person.
Adjusted	Boolean; indicates non-productive (0) and productive (1) audits.

Appendix B: Complex parameter, α, table for the decision tree

α	**Nbr of tree nodes**	**Learn error (no CV)**	**CV-learn error mean**	**CV-learn error std-dev**
0.150	1	1.0000	1.0000	0.04873
0.034	3	0.7000	0.7563	0.04398
0.031	4	0.6656	0.7375	0.04355
0.013	6	0.6031	0.6875	0.04235
0.006	8	0.5781	0.7063	0.04281
0.005	10	0.5656	0.6969	0.04258
0.005	13	0.5500	0.7063	0.04281
0.004	15	0.5406	0.7125	0.04296
0.003	20	0.5188	0.7156	0.04304
0.002	24	0.5062	0.7000	0.04266
0.002	27	0.5000	0.7031	0.04274
0.001	29	0.4969	0.7219	0.04318

Appendix C: Rules derived from the optimal decision tree (α = 0.01250 ≈ 0.013 in Appendix B)

```
Rule number: 13 [Adjusted=1 cover=11 (11/1333=1%) prob=1.00]
  Marital= Married
  Occupation= Cleaner, Clerical, Farming, Machinist, Repair, Service, Transport
  Deductions>= 1708

 Rule number: 15 [Adjusted=1 cover=176 (176/1333=13%) prob=0.79]
  Marital= Married
  Occupation= Executive, Professional, Protective, Sales, Support
  Education= Associate, Bachelor,  Doctorate,Master, Professional

 Rule number: 29 [Adjusted=1 cover=44 (44/1333=3%) prob=0.66]
  Marital= Married
  Occupation= Executive, Professional, Protective, Sales,Support
  Education=College, HSgrad, Vocational, Yr10, Yr11, Yr5t6, Yr7t8, Yr9
  Age=Mature

 Rule number: 28 [Adjusted=0 cover=76 (76/1333=6%) prob=0.37]
  Marital= Married
  Occupation= Executive, Professional, Protective, Sales, Support
  Education= College, HSgrad, Vocational, Yr10, Yr11, Yr5t6, Yr7t8, Yr9
  Age= Middle, Prime, Senior

 Rule number: 12 [Adjusted=0 cover=301 (301/1333=23%) prob=0.24]
  Marital= Married
  Occupation= Cleaner, Clerical, Farming, Machinist, Repair, Service, Transport
  Deductions< 1708

 Rule number: 2 [Adjusted=0 cover=725 (725/1333=54%) prob=0.06]
  Marital= Absent, Divorced, Married-spouse-absent, Unmarried, Widowed
```