Homework 5

Practice of Decision Trees

Description

The audit dataset is an artificially constructed dataset that has some of the characteristics of a true financial audit dataset for modelling productive and non-productive audits of a person's financial statement. 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 audit dataset is used to illustrate binary classification. The target variable is identified as TARGET_Adjusted.

The dataset is quite small, consisting of just 2000 entities. Its primary purpose is to illustrate modelling in Rattle, so a minimally sized dataset is suitable.

Variables

ID This is a unique identifier for each person.

Age The age.

Employment The type of employment.

Education The highest level of education.

Marital Current marital status.

Occupation The type of occupation.

Income The amount of income declared.

Gender The persons gender.

Deductions Total amount of expenses that a person claims in their financial statement.

Hours The average hours worked on a weekly basis.

Accounts The main country in which the person has most of their money banked.

Adjustment This variable records the 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 This is a numeric field of class integer, but limited to 0 and 1, indicating non-productive and productive audits, respectively. Productive audits are those that result in an adjustment being made to a client's financial statement.

Work to do:

1. Read the Audit.xlsx file and convert it to the csv extension.
2. The goal is to use a decision tree to predict the binary “Adjusted” variable, whether the individuals had made a correct financial statement or not. Decide which predictors you would use and eventually preprocess these variables.
3. Select the 1/3 of the last observations as test data.
4. Obtain the decision tree to predict whether the variable “Adjusted” on the training data. Decide the cutoff value for taking the decision.
5. Plot the importance of variables in the prediction.
6. Compute the accuracy, precision, recall and AUC on the test individuals.
7. Perform a Random Forest on the same data.