**Assignment #8: Decision Tree**

|  |
| --- |
| **Submission Instructions**   * **Submit the following four files** through the submission form:  1. The completed, working **R script** that produced the analysis with the complexity factor set to 0.05 in Part 1. 2. The output file **“DecisionTreeOutput.txt”** and **“TreeOutput.pdf”** for the analysis with the complexity factor set to 0.05 in Part 1. 3. The completed **answer sheet** (for Q1-Q10 in Part 1 and Q11-Q13 in Part 2).   **Evaluation**  Your submission will be graded based on the correctness of the completed answer sheet, with other files as supporting documents. |

**Part 1. Decision Tree Analysis in R**

**Before you start**

For this assignment, you’ll be working with the **BankLoan.csv** file and the **dTree.r** script (which we used in ICA #12). The BankLoan.csv file has data about 600 customers that received personal loans from a bank. The president of the bank wants to predict how likely a future customer is to pay back their loan so she can make better loan approval decisions.

The data file contains the following variables:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| **ID** | Customer identification number |
| **age** | The age of the customer, in years |
| **sex** | The gender of the customer |
| **region** | The type of area where the customer lives (INNER\_CITY, TOWN, SUBURBAN, RURAL) |
| **income** | Customer’s yearly income in dollars |
| **married** | Whether the customer is married |
| **children** | How many children the customer has |
| **car** | Whether the customer has a car |
| **save\_act** | Whether the customer has ever had a savings account with SchuffBank! |
| **current\_act** | Whether the customer has an active account with SchuffBank! |
| **mortgage** | Whether the customer has a mortgage |
| **payback** | Whether the customer paid back their loan (0 = no, 1 = yes)  **NOTE: payback** is the outcome variable we are interested in here. It describes a categorical event (0 = no, 1 = yes). |

**Guidelines:**

1. You’ll need to modify the script with the following information to perform the analysis:

* Set the input filename to the bank’s dataset (i.e., BankLoan.csv).
* Set the training partition (using TRAINING\_PART) to 60% (0.60) of the data set.
* Set the minimum split (using MINIMUMSPLIT) to 25.
* Set the complexity factor (using COMPLEXITYFACTOR) to 0.005.
* Make sure the outcome column setting is correct for your data set (using OUTCOME\_COL).
* You will need to modify the model to reflect the data set. This requires editing lines 82- 84 of the dTree.r script. Make sure you choose the correct outcome variable and exclude the variables that are inappropriate for the analysis. (HINT: ID is irrelevant to the analysis.)

1. Once you finish modifying the script, you can set the working directory and run the script.
2. **Based on your script output, answer Questions 1-6 in the answer sheet at the end of this document:**  
   *(NOTE: When asked “how likely…” cite the percentage!)*
3. **Now change the complexity factor from 0.005 to 0.05 and re-run the script. Using the new tree, answer Questions 7-10 in the answer sheet at the end of this document.**

|  |
| --- |
| **NOTE: Scientific Notations in R**  In the output, you may see numbers like-1.8e-04 or 31e+5.  The "e" is a symbol for base-10 scientific notation. The "e" stands for ×10exponent.  So -1.8e-04 means −1.8×10−4. In fixed-point notation that would be -0.00018.  Similarly, 31e+5 means 31×105. In fixed-point notation that would be 3,100,000. |

**Part 2. Compute and Evaluate Decision Trees**

Consider the following based on a different, hypothetical data set.

**Question 11.** (write your answer in the answer sheet)

Suppose we run the decision tree algorithm and get a decision tree (called Tree #1): compute the correct classification rate based on the following confusion matrix (*Compute it with a calculator. No need to use R/RStudio*):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Predicted outcome:** | |  |
|  |  | 1 | 0 |  |
| **Observed outcome:** | 1 | 510 | 220 |  |
| 0 | 240 | 1030 | Total: 2000 |

Table 1. Confusion Matrix (Tree #1)

**Question 12.** (write your answer in the answer sheet)

Suppose we re-run the decision tree algorithm and get another decision tree (called Tree #2): compute the correct classification rate based on the following confusion matrix (Compute it *by hand. No need to use R/RStudio*):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Predicted outcome:** | |  |
|  |  | 1 | 0 |  |
| **Observed outcome:** | 1 | 820 | 120 |  |
| 0 | 380 | 680 | Total: 2000 |

Table 2. Confusion Matrix (Tree #2)

**Question 13.** (write your answer in the answer sheet)

Which decision tree (Tree #1 versus Tree #2) has higher classification accuracy?