**Decision Tree Analysis in R**

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

**Before you start**Make sure you are using the most recent R/Rstudio version!

For this assignment, you’ll be working with the **BankLoan.csv** file and the **dTree.r** script (which we used in ICA #11). 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 **65% (0.65)** 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 data set than what you have done so far in this assignment.

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

Suppose we run the decision tree algorithm and get a decision tree (called it Tree #1): 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 | 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 it 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?

*Fill in the answer sheet below.*

|  |  |  |
| --- | --- | --- |
|  | **Question** | **Answer** |
| **Part 1. Decision Tree in R**  **(Complexity factor = 0.005)** | | |
| 1 | How often will this tree make a correct prediction (include decimals)? Provide your answer for both the training set and the validation set. | Training set **88.61%**  Validation set **79.17%** |
| 2 | How likely is a customer to pay back their loan if they have one child, at age 38 and make $35,000 per year?  *(NOTE: When asked “how likely…” cite the percentage!)* | **96%** likely |
| 3 | How likely is a customer to pay back their loan if they are married, at age 40, make $45,000 per year, have no children, have no mortgage and had saving account? | **9%** likely |
| 4 | How likely is a customer to pay back their loan if they make $52,000 per year, have three children, at age 41, and have saving account? | **87%** likely |
| 5 | Describe the profile of the least likely customer to successfully repay their loan. | Someone with an income of less than $15,000 and one child is **7%** likely to successfully repay their loan. |
| 6 | Describe the profile of the most likely customer to successfully repay their loan. | Someone who is unmarried with no children and no mortgage with an income less than $48,000 is **97%** likely to successfully repay their loan. |
| **(Complexity factor = 0.05)** | | |
| 7 | How often will this new tree make a correct prediction (include decimals)? Provide your answer for both the training set and the validation set. | Training set **85.83%**  Validation set **75.42%** |
| 8 | Is this model better or worse than the first model at predicting who will repay their loan? Explain how changing the complexity factor affected the tree using **no more than two sentences.** | This model is **worse** than the first model at predicting who will repay their loan. Increasing the complexity factor from 0.005 to 0.05 makes the tree **easier** to read (more simple), but **less** accurate. |
| 9 | How likely is a customer to pay back their loan if they have two children and make $35,000 per year? | **69%** likely |
| 10 | Does marriage increase or decrease the likelihood that a customer will pay back their loan? | Marriage **decreases** the likelihood that a customer will pay back their loan. |
| **Part 2 Compute and Evaluate Decision Trees** | | |
| 11 | What is the correct classification rate for Tree #1? | 0.77 or **77%** |
| 12 | What is the correct classification rate for Tree #2? | 0.75 or **75%** |
| 13 | Which decision tree (Tree #1 versus Tree #2) has higher classification accuracy? | **Tree #1** |