MIS2502: Data Analytics
Assignment: Decision Tree Induction Using R

For this assignment, you'll be working with the BankLoan.csv file and the dTree.r script. This 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 fields:

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 cars
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)

You'll need to modify the script with the following information to perform the analysis:

- Set the input filename to the bank's dataset.
- Set the training partition to 50% of the data set.
- Set the minimum split to 25.
- Set the complexity factor to 0.005.
- Make sure the outcome column setting is correct for your data set.
- You will need to modify the model to reflect the data set. This requires editing lines 76, 77, and 78 of the dTree.r script. Make sure you choose the correct outcome variable and you exclude the variables which are inappropriate for the analysis.

Answer the following questions (complete the worksheet at the end of this document):
(NOTE: When asked "how likely…" cite the percentage!)

1) How often will this tree make a correct prediction (include decimals)?
2) How likely is a customer to pay back their loan if they have one child and make $35,000 per year?
3) How likely is a customer to pay back their loan if they are married, make $45,000 per year, have no children, and no mortgage?
4) How likely is a customer to pay back their loan if they make $83,000 per year and have no children?
5) Describe the profile of the least likely customer to successfully repay their loan.
6) Describe the profile of the most likely customer to successfully repay their loan.

Now change the complexity factor from 0.005 to 0.05 and re-run the script. Using the new tree, answer the following questions:

7) How many leaf nodes are in the new tree?
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 **no more than two sentences**.
9) How likely is a customer to pay back their loan if they have one child and make $35,000 per year?
10) Does marriage increase or decrease the likelihood that a customer will pay back their loan?

What to submit:

Send a single email to your instructor with the following attachments:

- The completed, working R script that produced the analysis with the complexity factor set to 0.05.
- The output file "DecisionTreeOutput.txt" and "TreeOutput.pdf" for the analysis with the complexity factor set to 0.05.
- The completed worksheet provided on the last page.

Compute and Evaluate Chi-Squared Statistics

Consider the following based on a different data set than what you have done so far in this assignment.

11) Compute the Chi-Squared statistic for the following potential split variables:
(Note that you'll need to construct the "expected" distributions for each variable to come up with the Chi-Squared statistic!)

	Observed for PromSpend (total dollars spent at store)		
	<50	>=50	
Buy	520	730	1250
No Buy	480	770	1250
	1000	1500	2500

	Observed for PromTime (months as loyalty card member)		
	<6	>=6	
Buy	370	880	1250
No Buy	630	620	1250
	1000	1500	2500

12) Which variable is a stronger differentiator (PromSpend or PromTime) with regard to whether a consumer buys organics?

Answer Sheet for Assignment: Decision Tree Induction Using R

Name _____

Fill in the worksheet below with the answers to the questions on page 2 of the assignment:

Question	Answer
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	Answer for PromTime: Answer for PromSpend:
12	