**INST 354**

**Exercise 4**

**Deadline: 12/13/20, 11:59 PM**

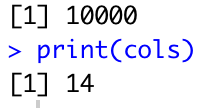
Please submit your work on Canvas. As a reminder, **this is individual work**. Submit your own work, in your own files. Name your files “your\_last\_name\_exercise\_4”. For example, “antoun\_exercise\_4.doc”. Please submit your answers in a Word document.

For this assignment, use the **bankchurn.csv** dataset on the assignment page. The dataset contains information about bank customers as well as whether they ended up closing their accounts or not. You need to be able to predict which customers will end up closing their accounts. Load the dataset into R and answer the 10 questions below (each worth 1 point).

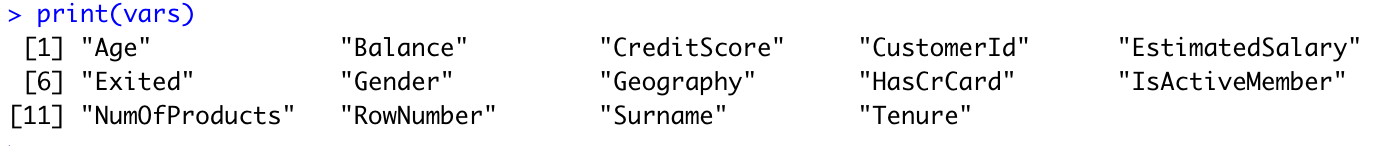
1. Suppose you are trying to design practical ways to retain bank customers (perhaps by providing special offers?). However, you are unsure of what to do to make sure that the programs will be successful. Describe how you might use the **bankchurn** data to help make your decision.

I will use the bankchurn data in order to identify specific variables that may help alleviate the issue. To be more precise, we must use subjective and value components to specify a set of alternatives that may be difficult in order to map one one’s general concerns or values about the special offers that are given from the bank. In addition, this will allow us to use good metrics to indicate the specific concepts or ideas that work. The bankchurn data will allow us to understand whether a customer ends up closing their account or not, therefore, the bank can send special offers to those specific customers.

1. How many rows and columns does the **bankchurn** dataset have?



1. What are the variables in the **bankchurn** dataset? Which ones might be relevant in trying to determine whether a customer ends up closing their account or not? Which variables, if any, have NA values in them?



In trying to determine whether a customer ends up closing their account you must use balance, exited, isactivemember, and tenure. There were no variables with NA values in them.

1. account and 0 means they have not). Since it's been coded as numbers, we need to change it to a categorical variable to make sure the `R` recognizes it as such. Show the code for how you did this.

dfBankChurn$Exited = factor(dfBankChurn$Exited)

1. What is the total number of people who have closed their account? What is the total number of people who have not closed their account? What is the average age of customers? What is their average credit score? What is the average number of years for which customers have been with the bank?

The total number of people who have closed their accounts is 2037.

The total number of people who have not closed their account is 7963.

The average age of customers is 38.9218.

The average credit score is 650.5288

The average number of years for which customers have been with the bank is 5.0128

1. Make sure you do the appropriate data cleaning and split the data into 25% test set and 75% train set. Show the code for how you did this.

numTest = floor(nrow(dfBankChurn) \* 0.25)

testRows = sample(1:nrow(dfBankChurn), numTest)

testDfBankChurn = dfBankChurn[testRows,]

trainDfBankChurn = dfBankChurn[-testRows,]

1. Run a tree model to predict ‘Exited’ (closed account vs retained) using credit score, age, and tenure as the features. Show your code and tree plot.

trainDfBankChurn$Exited = factor(trainDfBankChurn$Exited)

treemod = rpart(Exited ~ CreditScore + Age + Tenure,

data = trainDfBankChurn,

method = 'class',

control = rpart.control(minsplit = 25))

fancyRpartPlot(treemod, sub = "")

A picture containing text, clock

Description automatically generated

1. Find the prediction scores and set the top 20% as the threshold for a "successful" prediction. Display the confusion matrix. (Hint: In this matrix, you will have two columns and two rows. The two columns will be either your actual data or predicted data – be sure

to label what they are. Similarly, the two rows will be either your actual data or predicted data – be sure to label what they are.)

Table

Description automatically generated

1. Report the precision and the recall. Interpret what they mean in context.

Graphical user interface, text, application

Description automatically generated

The precision measures the percentage of our results that are relevant which is 49% and the recall rate allows us to understand the percentage of total relevant results correctly classified by our algorithm. This allows us to understand how accurate our information is which reveals a 50-50 shot.

1. Would you feel confident using this predictive model to make business decisions? Why or why not? (2-3 sentences)

I personally would feel confident using this predictive model to make business decisions as it allows us to understand specific variables and the statistics associated with them. With these statistics, we can analyze who to give special offers to base on tenure or credit score of the people. I would use this model to make sure more people join the bank, rather than leave it.