AD699: Data Mining for Business Analytics
Individual Assignment #4
Spring 2021
Due Date: 05APR, 11:59 p.m.

You will submit two files:

(1) **a PDF with your write-up**, along with
(2) **the script you used** to generate your results.

Task 1: Classification Tree

1. Bring the dataset *telecom_users.csv* into your R environment.

2. Convert the variable *Churn* into a factor. This will be your outcome variable for everything prior to Step 12 (Bug Bounty: Shang)

3. There are two variables in the dataset that will have no predictive value in the model. Using any method that you prefer to use, determine what they are, and then delete them.
    a. In just a sentence or two, how did you determine that these were the two useless variables?

4. Using your assigned seed value (from Assignment 2), partition your data into training (60%) and validation (40%) sets. Show the step(s) that you used to do this.

5. It's time to get weird! Build the LARGEST tree you can possibly make. Call that *model1*.
    a. Determine model1's accuracy against the training set. What is it?
    b. Determine model1's accuracy against the validation set. What is it?

6. Now build a very small tree. Be as crazy as you want to be. Just make sure this tree is small. Call it *model2*.
    a. Use rpart.plot to see your tree.
    b. Determine model2's accuracy against the training set. What is it?
    c. Determine model2's accuracy against the validation set. What is it?

7. Now build an optimally-sized tree. Call this one *model3*. Demonstrate the process that you used for finding the optimal tree size.

a. Using rpart.plot, show your tree model (you may wish to play around with the 'type' and 'extra' parameters of rpart.plot to adjust the way the tree looks).
b. Determine model3's accuracy against the training set. What is it?
c. Determine model3's accuracy against the validation set. What is it?

8. What general trend did you notice with accuracy as you moved from model1 to model2 to model 3? Describe this in a paragraph of about 3-5 sentences, being sure to include some thoughts about ***why*** the accuracy changed the way that it did as you tried these different models.

9. Describe the split that's created at your tree's root node (what variable did it split on, and what rule did it use?). Why is the root node significant?

10. Describe any one rule that your tree generates about whether a customer will churn. To describe a rule, just trace any path along your tree from the root node to a terminal node.

11. Pick ANY node, ANYWHERE on your tree, and calculate its Gini impurity. Demonstrate the calculation in R with some basic console math -- no fancy functions here! (hint: ?rpart.plot might be worth exploring in order to see the numbers and types of records in some node).

12. Smart bins/Dumb bins
    a. Make a copy of your dataset. Then, take one of the numeric outcome variables, and bin it into groups that have similar numbers of records. Then, partition the data into a 60/40 training/validation split. (Bug Bounty: Kelly)
    b. Now, build a new tree model, using that newly-binned variable as the outcome, and all of the other variables you used from model1/2/3 as your inputs. Don't worry about optimal sizing here -- just use all the defaults in rpart for this.
    c. What was your model's accuracy against the validation set?
    d. Now, with a new copy of the dataset, re-bin again, to convert that same numeric outcome variable into a factor with four bins. But this time, do the binning in a **terrible** way. Be uneven, be imbalanced...be awful.
    e. Re-partition the data, and build another model (like the one you made earlier in this step) but this time with the awful outcome variable.
    f. What was your model's accuracy against the validation set?
    g. Finally, compare the accuracy of the models you made in this step. What can you say about the comparative accuracy of these models, as well as their comparative quality? What sort of takeaway/conclusion can you draw from this exercise? (A couple sentences here will be enough to express this).

Task 2: Association rules

For this portion of the assignment, we will be using data from Groceries, a dataset that can be found with the *arules* package. Each row in the file represents one buyer's purchases. This link provides some helpful templated examples for generating association rules:

http://r-statistics.co/Association-Mining-With-R.html

1. Describe "Groceries" by answering following questions:
   - What is the class of "Groceries"?
   - How many rows and columns does Groceries contain?

2. Generate an item frequency barplot for the grocery items with support rate greater than 5%. Include a screenshot of your results, along with the code you used to do this.

3. Now, create a subset of rules that contain **your grocery item** (you can find your item in the spreadsheet in Blackboard). Select any **one** rule with your item on the left-hand side, and any **one** rule with your item on the right-hand side, and explain them in the way you would explain them to your roommate (I'm assuming your roommate is a smart person who is unfamiliar with data mining). *Remember, every rule has four components: support, coverage, confidence, and lift*.

   For each of your chosen rules (your grocery item on the left-hand side, and your grocery item on the right-hand side), include a screenshot of your rules, along with the code you used to generate the rules.

4. In a sentence or two, explain what meaning these rules might have for a store like Star Market. What could it do with this information?

5. Using the plot() function in the arulesViz package, generate a scatter plot of *any* three rules involving your grocery item. Include a screenshot of your plot, along with the code you used to generate the plot. Describe your results in a sentence or two.

6. Again using the plot() function in the arulesViz package, generate a plot for any three of your rules. This time, add two more arguments to the function: method="graph", engine="htmlwidget". What do you see now? Include a screenshot of your plot, along with the code you used to generate the plot. Describe your results in a sentence or two. In your answer, be sure to explain what the size of the circle, and shading of the circle,

indicate. If you're not sure, remember to be resourceful! Help menus, searches, and some trial-and-error could go a long way here.