Problem Set 4: LMS Module 8

YOUR NAME HERE

due INSERT DATE HERE

Structural stuff:
1. Insert a YAML header above (remember, the three dashes must be on line 1 in the .Rmd file). If you're not sure what the header looks like, there's a screenshot of one on the course website. Modify the header to have your name as the author and the due date for this assignment as the date. 2. Before you knit, save your .Rmd file as LastName_FirstName.Rmd.
3. You need to submit your .Rmd code file AND a knit file (upload both simultaneously to the course webpage; you can't upload them one-by-one). You will only receive full credit if you upload both files (note: you get 20 points JUST for uploading both). 4. Below I have set up the file for you with the libraries you'll need. I have also inserted code chunks for you.
5. I expect that the .Rmd file you submit will run cleanly, and that the knit file won't contain any errors (LOOK at the knit file after you create it - if questions/text are running into each other, if you see error messages, etc., you're not done). 6. You can use comments to tell me what you are doing either in text or in code chunks, but remove "old" code that didn't run/work.

"When life gives you lemons, don't make lemonade. Make life take the lemons back! Get mad!" - Cave Johnson

This assignment includes questions from Module 8 (classification). These exercises will require you to load the "lemon" .csv file that includes data from Carvana/Kaggle on cars and whether they are considered "lemons" (or "bad buys"). You can find more information about the data here:.

Research Topic: For this exercise, we are taking on the role of advisor at a local organization that supports young adults as they transition out of high school into postsecondary education and/or the workforce. As a part of our guidance to them on making smart financial decisions, we want to see if there are ways to help them prevent financial pitfalls, such as purchasing a "lemon" of a car. Our dependent variable for this analysis will be whether or not a vehicle is a "bad buy" based on a set of predictors from the dataset we have available to us. Make sure you keep this goal in mind when you are analyzing and interpreting findings from this data.

Lemon Dataset Codebook can be found here: under "Carvana_Data_Dictionary.txt"

1. Import/load the dataset here (take note of what type of file this is to ensure you use the right code to read it in). (4pts)
2. Calculate the proportion of lemons in the dataset using the `IsBadBuy` variable. (4pts)

2a. What proportion of cars in this data are lemons? (2pts)

3. Plot the dependent variable using the appropriate univariate plot (4pts).
4. Calculate the proportion of lemons by our first independent variable of interest: `Make` (4pts).

4a. Which make of car has the highest proportion of lemons and what is that proportion? (2pts)

4b. Which make of car has the lowest proportion of lemons and what is that proportion? (2pts)

5. Calculate the proportion of lemons by second independent variable of interest: vehicle age (`VehicleAge`) (4pts)

5a. Do you see a potential association between vehicle age and the probability of a car being a lemon? If yes, what does that association tell you? If no, tell me why there's no association (how do you know that from the data)? (2pts)

6. Let's examine the proportion of lemons by a third independent variable of interest: odometer reading (`VehOdo`). Because this variable is continuous, let's break it into quintiles, or 5 groups (hint: use the ntiles code from the Module 2 .Rmd file) and then calculate conditional mean/proportions for those 5 groups. (6pts)

6a. Do you see a potential association between odometer reading and the probability of a car being a lemon? If yes, what does that association tell you? If no, tell me why there's no likely association (how do you know that from the data)? (2pts)

7. Before running a logistic regression/classification model, you need to make your outcome variable into a factor variable. Make sure you convert it to a factor and save it to your dataset. (2pts)

8. Use set.seed(5634) and split the dataset to create testing/training datasets. (2pts)

9. Now, let's predict the probability of being a lemon using a logistic regression/classification model and the three independent variables we examined above (Make, Vehicle Age, and Odometer Reading). Run this model using your training data. In this chunk you should set the model, set up the recipe, define the model, put the workflow together, fit the results, and then present your tidy table (these are all the steps we went through together in the code - nothing new here). NOTE: PLEASE PUT `MAKE` as the LAST variable in your regression equation (so that the estimated for Vehicle Age and Odometer Reading are first in your results table). (10pts)

9a. Provide a sentence interpreting the intercept of your model. (2pts)

9b. Provide a sentence interpreting the slope of `VehicleAge` (remember, you can exponentiate your estimate to obtain an odds ratio before interpreting). (2pts)

9c. Provide a sentence interpreting the slope of `VehOdo` (remember, you can exponentiate your estimate to obtain an odds ratio before interpreting). (2pts)

9d. Provide a sentence interpreting the slope for the "Nissan" category from your `Make` variable (THINK about what the referent group is for `Make` in this analysis). (2pts)

10. Use your logit model results to make predictions in your testing dataset and create a confusion matrix. (4pts)

10a. How many "NOs" did you predict correctly? How many "YESs" did you predict correctly? (2pts)

11. Calculate the accuracy, sensitivity and specificity of your model. (4pts)

11a. Provide a sentence interpreting the accuracy of this model. (2pts)

11b. Provide a sentence interpreting the sensitivity of this model. (2pts)

11c. Provide a sentence interpreting the specificity of this model. (2pts)

12. Calculate the AUC for the predictions from the ROC based on the logit model. (4pts)

12a. Provide one sentence interpreting this roc_auc. (2pts)