Problem Set 5: LMS Modules 9 & 10

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.

This assignment includes questions from Modules 9 (classification plots) and 10 (cross validation). These exercises will require you to load the "Lemons (car=bad buy) dataset" from the course website that includes data in a .csv file from Carvana/Kaggle on cars and whether they are considered "lemons" (or "bad buys").

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

Research Topic: For this exercise, we are expanding on our analysis from Problem Set 4 by producing some visual depictions of our data for a presentation to key stakeholders at our local organization. We are also going to cross validate our model to ensure that what we have found will hold up when using it to make predictions in out-of-sample data.

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. Plot a bar graph that shows the mean Vehicle Age (`VehicleAge`) age for each of our possible outcome (`IsBadBuy`) values (i.e., 0, 1). Your plot MUST include axis labels and a title. Here are a few hints to help: Hint 1 - We want the mean age for each level of our dependent variable, so think about what you'll group_by in this case. Hint 2 - While typically we always plot our DV on the y-axis, in this case, you'll want `IsBadBuy` to be plotted on your x-axis in your plot. Hint 3 - add as.factor to `IsBadBuy` (i.e., as.factor(IsBadBuy)) in your plot code so that you only get data axis values of 0 and 1 (rather than nonsensical values like 0.5). (10pts)

2a. What was the difference in the mean age between cars that weren't lemons and those that were? Did bad buys tend to be older or younger? (4pts)

3. Plot a bar graph that shows the mean odometer reading (`VehOdo`) for each of our possible outcome (`IsBadBuy`) values (i.e., 0, 1). Your plot MUST include axis labels and a title (also, you may want to see the hints provided in question 2 above). (10pts)

3a. What was the difference in the mean odometer reading between cars that weren't lemons and those that were? Did bad buys tend to have higher or lower mileage? (4pts)

4. Run a logistic regression/classification model using the two IVs listed above as predictors (`VehicleAge`, `VehOdo`) using the FULL dataset (don't split into testing/training here). Remember to make your IsBadBuy variable a factor if you haven't already and be sure you show your regression summary table. (10pts)

Remember that the output of logistic regression is in log odds. You may want to convert to odds ratios/percent change for the interpretations below.

4a. Provide a sentence interpreting the intercept for the model. (2pts)

4b. Provide a sentence interpreting the estimate for age. (2pts)

4c. Provide a sentence interpreting the estimate for odometer. (2pts)

5. Generate a heat map for the probability of being a bad buy by vehicle age and vehicle odometer reading. Make each group into quintiles before doing this. (10pts)

5a. What does your heat map show you about the predicted probability of a car being a bad buy based on these two independent variables? (2pts)

6. Instead of using training/testing splits to validate your model, we are going to run a monte carlo cross-validation for this classification model. Specify 50 repeated samples for your analysis. (10pts)

6a. Provide a sentence interpreting this model's sensitivity. (2pts)

6b. Provide a sentence interpreting this model's specificity. (2pts)

6c. Provide a sentence interpreting this model's roc_auc. (2pts)

6d. What are your conclusions about the model overall? How might you improve it (or, if you don't think it needs improvement, tell me why)? (4pts)