Assignment 6 - Classification Plots & Cross Validation

YOUR NAME HERE

12/1/21

For this assignment, you'll be using the lemons dataset (again). On the last assignment, your job was to predict which cars were most likely to be lemons. Now, you'll be plotting your results. You'll want to read in the lemons dataset from the course website. As always, please put any sentences for your question answers OUTSIDE of code chunks. For this analysis our outcome is still whether or not a car is a "bad buy" or "lemon." We're interested in the following two continuous independent variables: VehicleAge, and VehOdo (i.e., mileage).

1. Plot a bar graph that shows the mean age for each of our possible IsBadBuy values (i.e., 0, 1). Make sure you have axes labeled 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 your x variable in your plot. Hint 3 - add as.factor to IsBadBuy in your plot code so that you only get data axis labels for 0 and 1 (i.e., as.factor(IsBadBuy)).

1a. 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?

2. Plot a bar graph that shows the mean odometer reading for each of our possible IsBadBuy values (i.e., 0, 1). Make sure you have axes labeled and a title. Make sure you have axes labeled and a title (see the hints from Question 1).

2a. 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?

3. Run a logistic regression using the two IVs listed above as predictors (Age, Odometer). Remember to make your IsBadBuy variable a factor if you haven't already! Make sure you show your regression summary table.

3a. Provide a sentence interpreting the estimate for age (Hint: use the logistic regression powerpoint from the course website).

3a. Provide a sentence interpreting the estimate for odometer.

4. 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.

4a. What does your heat map show you about the predicted probability of a car being a bad buy based on these two independent variables?

5. Instead of using training/testing splits to validate your model, run a monte carlo cross-validation for your classification model. Specify 50 repeated samples for your analysis.

5a. What did the results from your cross-validation show you about your model?