Problem Set 4: LMS Module 8

YOUR NAME HERE

due INSERT DATE HERE

Structural stuff:
Same as all previous assignment :)

This assignment includes questions from Module 8 (classification). These exercises will require you to load the "cvdd.rdata" file from the course website that includes excerpted data from The Framingham Study on cardiovascular disease.

The National Heart, Lung, and Blood Institute of the National Institutes of Health developed a longitudinal, epidemiology-focused dataset using the Framingham Heart Study. The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease. This dataset contains three clinic examinations and 20 year follow-up data on a large subset of the original Framingham cohort participants.

Research Topic: For this exercise, we are interested in understanding which risk factors predict cardiovascular disease or death (CVDD). We will examine a set of categorical and continuous predictors and interpret their predictive value in understanding CVDD.

CVDD Dataset Codebook can be found here: https://rdrr.io/cran/riskCommunicator/man/cvdd.html

Additional variable we added:
* cvd_deathnum = numeric version of `cvd_dth` variable (you'll need this for numeric calculations)

1. Import the dataset here and examine it to begin to get familiar with the data. (4pts)
2. Calculate the proportion of people who either had cardiovascular disease or died from cardiovascular disease (CVDD) in the dataset (you can use the factor version of the outcome: `cvd_dth` or the numeric version `cvd_deathnum`). This will be our dependent variable for subsequent analyses. (4pts)

2a. What proportion of people in this dataset had CVDD? (2pts)

3. Plot this dependent variable using the appropriate univariate plot (Hint: use the factor version of the DV for plotting). (6pts)
4. Calculate the proportion of people who had CVDD by our first independent variable of interest: Current smoking at exam - `CURSMOKE` (Hint: use the numeric version of the DV for calculations). (4pts)

4a. Which group had a higher proportion of people with CVDD? (2pt)

4b. Was this what you expected (explain why)? (2pt)

5. Calculate the proportion of CVDD by our second independent variable of interest: BMI category (`bmicat`) (Hint: use the numeric version of the DV for calculations). (4pts)

5a. Do you see a potential association between BMI categories and the probability of someone having CVDD? 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 CVDD by a third independent variable of interest: Age (`AGE`). 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. (4pts)

6a. Do you see a potential association between age and the probability of someone having CVDD? 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)

Now that we have examined three IVs, let's go ahead and build a logistic regression model.

7. Use set.seed(5634) and split the dataset to create testing/training datasets. (4pts)

8. Now, let's predict the probability of CVDD using a logistic regression/classification model and the three independent variables we examined above (`CURSMOKE`, `bmicat`, `AGE`). Run this model using your training data. In this chunk you should define the formula, define the model, set up the recipe, put the workflow together, fit the results, and then present your tidy table (these are all the steps we have gone through together - nothing new here). (Hints: be sure to use the factor version of your DV for your analysis). (10pts)
   IMPORTANT NOTES: Nothing needs to be log transformed here (so if you see "step_log" in your code, remove it!). However, we DO want to convert our categorical/non-numeric predictors to dummy variables. You should add this code when you define your recipe:
   `%>%step_dummy(all_nominal(), -all_outcomes())`

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

9b. Provide a sentence interpreting the slope of category 1 from your `CURSMOKE` variable (remember, you can exponentiate your estimate to obtain an odds ratio before interpreting and/or calculate percent change). (2pts)

9c. Provide a sentence interpreting the slope for category 3 from your `bmicat` variable (remember, you can exponentiate your estimate to obtain an odds ratio before interpreting and/or calculate percent change). (2pts)

9d. Provide a sentence interpreting the slope for your `AGE` variable (remember, you can exponentiate your estimate to obtain an odds ratio before interpreting and/or calculate percent change). (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 model's ROC_AUC. (2pts)

12a. Plot the roc_auc curve for this model. (2pts)

12b. Provide a sentence or two interpreting your roc_auc. (2pts)