Logistic Regression (Classification)

In this assignment you will complete a variety of tasks related to binary classification with logistic regression. The dataset that we will be using is related to criminal justice and deals specifically with parole violations.

Deliverable: All of your work for this assignment should be done in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.

Libraries: For this assignment you will need the following libraries: tidyverse, MASS, caret, ROCR

Before beginning the assignment tasks, you should read-in the data for the assignment into a data frame called parole. **Carefully** convert the male, race, state, crime, multiple.offenses, and violator variables to factors. Recode (rename) the factor levels of each of these variables according to the description of the variables provided in the ParoleData.txt file (located with the assignment on Canvas).

Task 1: Split the data into training and testing sets. Your training set should have 70% of the data. Use a random number (set.seed) of 12345.

Task 2: Our objective is to predict whether or not a parolee will violate his/her parole. In this task, use appropriate data visualizations and/or tables to identify which variables in the training set appear to be most predictive of the response variable "violator". Provide a brief explanation of your thought process. **Hint: When plotting two categorical variables against each other, consider using a barplot with geom_bar(position="fill"). See what you get when you try this**

Task 3: Identify the variable from Task 2 that appears to you to be most predictive of "violator". Create a logistic regression model using this variable to predict violator. Comment on the quality of the model.

Task 4: Using forward stepwise, backward stepwise, or by manually building a model, create the best model you can to predict "violator". Use only the training data set and use AIC to evaluate the "goodness" of the models. Comment on the quality of your final model. In particular, note which variables are significant and comment on how intuitive the model may (or may not) be.

Task 5: Create a logistic regression model using the training set to predict "violator" using the variables: state, multiple.offenses, and race. Comment on the quality of this model. Be sure to note which variables are significant.

Task 6: What is the predicted probability of parole violation of the two following parolees? Parolee1: Louisiana with multiple offenses and white race Parolee2: Kentucky with no multiple offenses and other race

Task 7: Develop an ROC curve and determine the probability threshold that best balances specificity and sensitivity (on the training set). **Hint: In the predict function, use type = "response" and do not use the [,2] that we used in the logistic regression threshold lecture. We only had to include that code in that lecture because we used k-fold cross validation**

Task 8: What is the accuracy, sensitivity, and specificity of the model on the training set given the cutoff from Task 7? What are the implications of incorrectly classifying a parolee?

Task 9: Identify a probability threshold (via trial-and-error) that best maximizes accuracy on the training set.

Task 10: Use your probability threshold from Task 9 to determine accuracy of the model on the testing set.