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.

Libraries: For this assignment you will need the following libraries: tidyverse, tidymodels, e1071, and 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). Take your time and double-check that you have correctly converted and renamed the variables listed above.

Question 1 There are 675 parolees in the dataset. How many of these parolees ended up violating parole? HINT: Examine the response variable "violator".

Question 2: 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. Be sure that the split is stratified by "violator".

Before proceeding, let's take a moment to talk about the ordering of the levels (categories) in the response variable. The command below shows us the levels of the response variable. We should expect them to be "No" and then "Yes" (in that order).

```
levels(train$violator)
```

Ordering is important when it comes to the categories of the response variable. We need the "positive" class (category) to be listed second. Here "Yes" is listed second. "Yes" is our "positive" class as we are interested in building models to detect parolees that violate parole rather than building models with the intent of identifying the parolees that do not violate parole. It seems like a small issue, but it's an important one. What do we do if the categories are in the incorrect order (this happens sometimes)? We can rearrange the factor levels to put the positive class second (last). The code below accomplishes this. If your levels are properly ordered already, it won't hurt to run this code. It's good to keep this code around in case you do need to reorder levels.

```
train = train %>% mutate(violator = fct_relevel(violator, c("No","Yes")))
levels(train$violator)
```

Question 3: 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 examine the relationship between each variable and the response variable "violator". Use your visualizations to answer the questions below.

True/False: The violation rate appears slightly higher among males than among females.

Question 4: True/False: The violation rate is considerably higher in Louisiana than in the other states.

Question 5: True/False: The violation rate appears slightly higher among parolees with shorter "max_sentence" values.

Question 6: Create a logistic regression model using the "state" variable to predict "violator".

Which state is the base level in the model summary?
A. KY
B. LA
C. VA
D. Other

Question 7 To two decimal places, what is the AIC of the model with "state" to predict "violator"?

Question 8 Create a logistic regression model using the training set to predict "violator" using the variables: "state", "multiple.offenses", and "race".

Which variables are significant in the resulting model (select all that are significant)?
A. state

B. multiple.offenses
C. race
D. None of the variables in the model are significant

Question 9: Use your model from Question 8 to determine the probability (to two decimal places) that the following parolee will violate parole: The parolee is in Louisiana, has multiple offenses, and is white.

Question 10: Continuing to use your model from Question 8, develop an ROC curve and determine the probability threshold that best balances specificity and sensitivity (on the training set). Be sure to be careful with the predict function syntax.

What is the value of this threshold (to four decimal places)?

Question 11: Continuing to use your model from Question 8, what is the model's accuracy (on the training set) given the cutoff from Question 10? Report the accuracy to three decimal places. HINT: Use the threshold value out to all of its reported decimal places to ensure that your answer matches the solution,

Question 12 Continuing to use the model from Question 8, what is the sensitivity of the model on the training set (to three decimal places)?

Question 13: For the model from Question 8, which probability threshold results in the best accuracy (on the training set)?
A. 0.2
B. 0.3
C. 0.4
D. 0.5

Question 14: Use your probability threshold from Question 13 to determine the accuracy of the model on the testing set (to three decimal places).