Stat 6021: Homework Set 12

1. For this question, we will revisit the `penguins` data set from the `palmerpenguins` package. The data set contains information regarding measurements of adult penguins near Palmer station, Antarctica. We will focus on using the four measurement variables (bill length, bill depth, flipper length, body mass) to model the gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression.

   When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame. Then, randomly split your data into a training and test set (80-20 split respectively). For reproducibility, use `set.seed(1)` while performing the split.

   From the last homework, you should have dropped flipper length from the model, while keeping bill length, bill depth, body mass, and species as predictors.

   (a) Validate your model on the test data by creating an ROC curve. What does your ROC curve tell you?

   (b) Find the AUC associated with your ROC curve. What does your AUC tell you?

   (c) Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?

   (d) Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

2. Please remember to complete the Module 9 to 12 Guided Question Set Participation Self- and Peer-Evaluation Questions. Complete via Qualtrics. Link provided in Canvas in the same place as where you found this PDF.