July 22

I. Open discussion
   A. Interpretative versus Machine languages
      1. Python and R are commonly classified as interpretive languages. Commands, objects and arguments are easier to understand, and the hard work is left for the interpreter
      2. Interpretative computing languages tend to be slower than machine programming languages (such as C, java etc…).
      3. Interpretative languages tend to be easier to learn and generally speaking more fun to use
      4. Both python and R have drawn in large audiences through their open source and open access structure
      5. Python typically excels at processing long lists, while also leveraging strong support from the open source community in terms of development.
      6. How is data science different from computational statistics? Computer science? Is it simply the combination of the two in the service of some discipline?Is it as simple as the difference between a scientist and engineer?

II. Problem Statement

III. Responses
   A. Boosted Trees
      1. What is a one-hot-encoded column and why might it be needed when transforming a feature? Are the source values continuous or discrete?
      2. What is a dense feature? For example, if you execute **example = dict(dftrain)** and then **tf.keras.layers.DenseFeatures(your_features)(your_object).numpy()**, how has the content of your data frame been transformed? Why might this be useful?
      3. Provide a histogram of the probabilities for the logistic regression as well as your boosted tree model. How do you interpret the two different models? Are their predictions essentially the same or is there some area where they are noticeable different. Plot the probability density function of the resulting probability predictions from the two models and use them to further illustrate your argument. Include the ROC plot and interpret it with regard to the proportion of true to false positive rates, as well as the area under the ROC curve. How does the measure of the AUC reflect upon the predictive power of your model?
   B. Boosted Trees continued (with model understanding)
      1. Upload your feature values contribution to predicted probability horizontal bar plot as well as your violin plot. Interpret and discuss the two plots. Which features appear to contribute the most to the predicted probability?
      2. Upload at least 2 feature importance plots. Which features are the most important in their contribution to your models predictive power?
      3. Stretch goal: Modify the visualization formula. Plot your output and provide an interpretation. Which estimator was more effective at reproducing the probability model that was used to generate the instance of data that was synthetically generated?