Aya Eid

Tempus Test Case

February 23, 2020

## Data:

This test case consisted of labelled data (binary classification of whether a patient has or does not have cancer), given a set of biomarkers and patient demographic information. This data was separated into three files that were linked through the patient ID. The goal was to build a model to best predict the classifier given a new patient. Input data consisted of over 15,000 features and 1960 patients. Additionally, there were only about 28% cancer patients. Thus, it was especially important to isolate the most important features to prevent overfitting, and ensure that the model put more weight on obtaining true positives since the data was slightly imbalanced.

## Model Description

Although I experimented with a few things, I ultimately landed on a Gradient Boosted (GBM) classifier as the final model for several reasons. The data set was slightly imbalanced in that it had four times as many negative targets as positives. The model required a little data clean-up, which included converting categorical inputs through one-hot encoding, filling in missing data points and removing features with too many missing observations. I started with a simple random forest classifier but due to the imbalance found that all predictions defaulted to false.

With the GBM, I used a grid search to ensure the learning rate and number of trees used to build the model were appropriate. Finally, I used another grid search to estimate the optimal tree depth to ensure it would not overfit the data and generalize well with new data. **Observing several important metrics, we see the model's accuracy was 0.89, area under the curve of 0.93, sensitivity of 0.88, specificity of 0.89, and ultimately a weighted F1-score of 0.90.** In another iteration of this model, I would attempt to retune the model parameters or change it to use extreme gradient boosting which would handle missing data more accurately.