**Module 3 Project Review: Classification**

I identified the data I needed by searching through the American Soccer Analysis website to find the pages that share Major League Soccer advanced player attacking metrics by year in the same format. I used Browser and Beautiful Soup to scrape the relevant web pages to get the attacking metrics data. I then turned the data into a DataFrame, which became easier once I realized that many entries had blank strings rather than zeroes or nulls.

I then explored the collinearity amongst the features and used PCA to include all of the features in a way that eliminated multicollinearity issues. I followed by pipelining SVM, Decision Tree, and Random Forest classifier models to find their respective explained variance amongst the data with default parameters. I then used a GridSearch to test for the best parameters for the SVM, Random Forest, and Adaboost classifiers and found that SVM with C = 1 and kernel = ‘linear’ resulted in the greatest explained variance. To validate the model, I repeated my methodology except I did not use a PCA, but rather I removed individual features that resulted in multicollinearity over .75. This second variation proved to explain a greater variance.

The results for the best classifier model (SVM, no PCA, C = 1, kernel = ‘linear’) was 56% accuracy in guessing the correct player position based on advanced attacking metrics. Since there were 7 position options, a random guess would have resulted in an accuracy of about 14%. It is a positive sign that the model is almost three times more effective. Some errors remained, however, a notable one being that goalkeepers were often mistaken for attackers or strikers. Accumulation and inclusion of defensive data would likely improve the model greatly. An improvement upon this promising model could allow teams to know the ideal fit of players they scout from foreign leagues and teams, which would bring great value to professional soccer clubs.