Jacob Kurstedt

**Final Report:**

**NFL Draft Combine**

**Problem Statement (How well a Player will do in the NFL):**

The NFL combine is a place where players can show what skills they have to improve their draft rank. Some players preform phenomenal in the drills boosting their draft rank and other players drop down due to poor performance. For example, the Tom Brady was projected to be a 4th round pick but since he had such a terrible combine, he dropped all the way down into the 6th round. All the analysts said he was going to be a bust however this was not the case and he turned out to be the greatest QB in history. Because of this I wanted to look more into seeing if the Combine had a relationship to whether a player would be great or a bust in the NFL based off defensive players.

**Data Wrangling/Cleaning:**

The first thing I had to do was find Data on players Combine stats. I used the website <https://nflcombineresults.com/> to find the data. I had to copy paste the tables they made into an excel file and then from there save the excel file as a csv in order to access the data. I wrangled the combine data from 2000-2010. I then had to clean up the tables. Next, I wrangled in the college records for the colleges the players went to and merged it with the combine stats data. From their I had to filter out all the defensive players from the data set which made the data set go from 3513 rows to 1620 rows. I then found a great website <https://www.pro-football-reference.com/years/2000/draft.htm> which provided me with data for all the drafted players between 2000 and 2010 and their career stats. I then cleaned the data up renaming columns and filtering out the offensive players. Then I merged this data set to my existing one and was left with my main data set which had the combine stats and career stats for all the defensive players between 2000-2010.

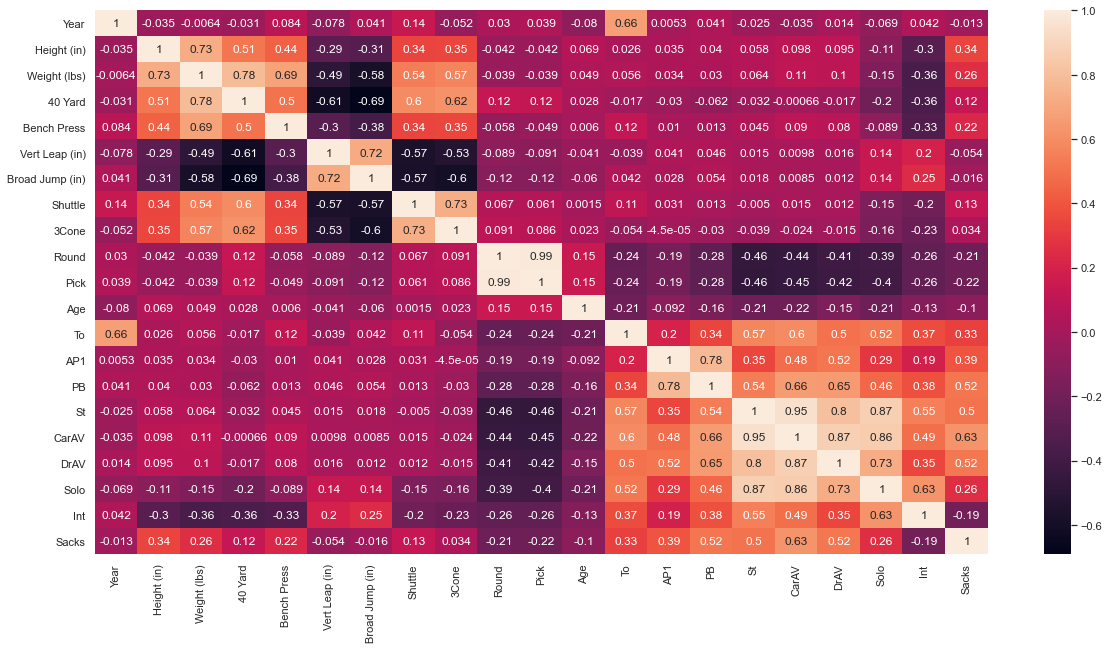
**Exploratory Data Analysis:**

The main parameters that I was looking at were all the draft Combine Skills tests which were:

* Height
* Weight
* 40 Yard Dash
* Bench Press
* Vertical Leap
* Broad Jump
* Shuttle run
* Three Cone drill

I explored each parameter individually to see the difference between players that were drafted and undrafted. For Bench Press, Vertical Leap, Shuttle Run and Three cone drill there was not really a difference in the mean between drafted and undrafted players by position which was very interesting. However, in the 40-yard dash and Broad Jump you could see that the players who were drafted preformed better in these tests.

I then looked at the correlations between the draft combine tests and the career stats. The career stats were Solo tackles, Interceptions, and Sacks. I saw that the correlations between them were very small meaning that there seemed to be no correlation between the draft combine tests and career stats, so I wanted to dig deeper.



**Preprocessing (How well a player will do in the NFL):**

During Preprocessing I made a feature called Pro-Bowl select which showed if a player made the pro bowl or not during their time in the NFL. I preformed many models on this feature and my accuracy scores for my models were not very good. I used Logistic Regression with and without PCA, Decision Trees, Random Forest, and Gradient Boosting with and without PCA. Each not getting that much accuracy when using Pro bowl select as my y variable and the draft combine test as my X variable. I also tried preforming some Unsupervised learning using KMeans and Agglomerative clustering. It seemed to do a pretty good job, but it was not answering the question that I was looking for. Since the scores were so low and could not give me a good insight into if the Combine Tests related to how well a defensive player would do in the NFL. I decided to change my problem statement.

**Problem Statement (Predict a Player getting drafted):**

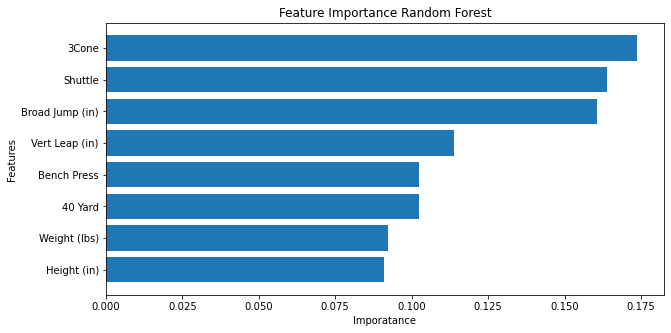
Since using the draft combine tests did not give me that good of insight I decided to try and predict which players would be drafted based of the combine scores they got. I already had the data I needed to answer this question so I did not have to wrangle anymore data or Explore the data anymore.

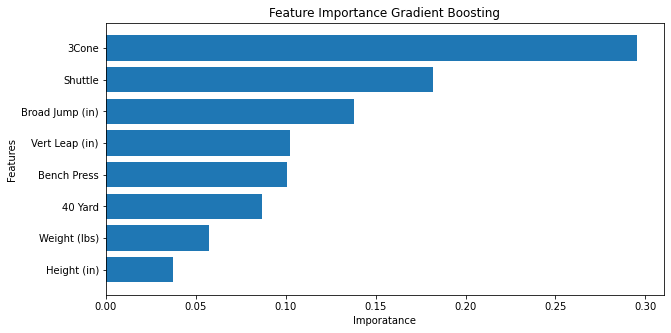
**Preprocessing (Predict a Player getting drafted):**

When preforming preprocessing on my new problem the only thing I had to change was my y variable to be if a player was drafted or undrafted. For each model I was getting a lot better score and same with Unsupervised learning as well I was getting better scores.

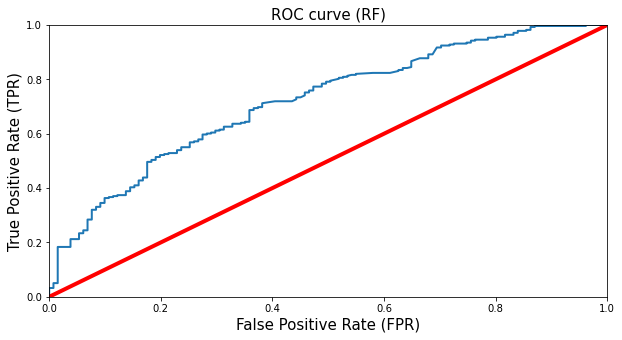
**Modeling/Selection:**

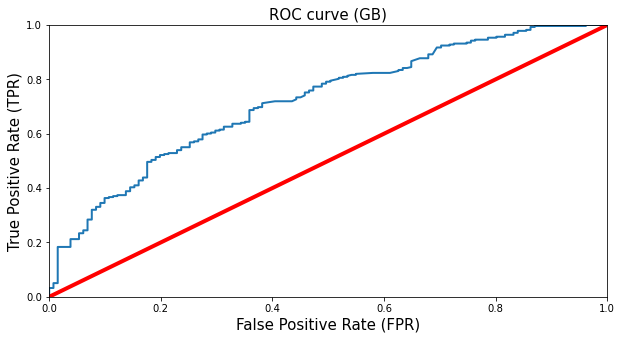
When choosing the best model, it came down to Random Forest and Gradient Boosting models. For both Random Forest and Gradient boosting I found out that each of them had the same exact order for feature importance with three cone drill being at the top. However, in Gradient Boosting the three-cone drill was by far the most important feature being .1 importance above the second important feature which was the shuttle run.





I then found the mean cross validation score for both the Random Forest and Gradient Boosting. For Random Forest, the mean score for train set was 0.68 and for the test set it was 0.66. For Gradient Boosting the mean score for the train set was 0.66 and the test set 0.67. After this I looked at the ROC curve and got the ROC-AUC score for each of the models. The Random Forest had a ROC- AUC score of 0.69 and the Gradient Boosting had a ROC-AUC score of 0.71. When looking at theses scores, I decided that the Gradient Boosting was the best choice. Both Models scored about the same when looking at the mean cross valuations scores however the ROC-AUC score for Gradient Boosting was 0.2 higher than the score for Random Forest. Thus I choose Gradient Boosting as the best Model.



****

**Conclusion:**

In the end Gradient Boosting was the best model to choose for predicting whether a player would be drafted or undrafted. It had the best ROC-AUC score and outperformed all the other models.

What I found very interesting though was that the most important features when predicting if a player was drafted was the Three Cone Drill, Shuttle Run, and Broad Jump. The reason I think this is because in football you have to be quick laterally and be able to change direction quickly which the three cone drill and shuttle run are testing. While the Broad jump test explosiveness which is very key to being a football player.

In all, being able to predict whether a player would be drafted or undrafted can help out many NFL teams since it could help them decide which player to choose over another even though they may have played in different conferences if their combine scores are relatively the same one player could be a potential diamond in the rough just like Tom Brady.

**Moving Forward:**

In the future when working on this project again I would extend my data to be the past 20 years. I will also perform the same things I did with Offensive players to see if the models perform better. I will also try to look for college data to see how they performed in college and what conference they were in to see if which conferences produced the most players in the NFL. Also I would look more in depth at the precision and recall scores since it would be better for a team to draft a player who would have gone undrafted otherwise. Lastly, when modeling I will perform PCA on my model selections to see if they performed better than the models I chose and create better graphs for more visualization.