ORIE 4741 Data Analysis Report

The Most Important Features to Consider in NFL Draft Prospects

Kyle Bleustein (ksb224), Kira Solberg (kms389)

Abstract

Every year more than 200 players are drafted into the National Football League (NFL) and even more players participate in the NFL combine. During the NFL combine, each player participates in different skill tests and their statistics are recorded and used as a source of comparison with other players. NFL general managers use the NFL combine and other information to create a draft strategy for drafting new players. With the surplus of information, it is hard to tell which tests and player characteristics are the most significant in determining their draft position. We are interested in exploring what features are most important for the NFL combine in determining when a player will be drafted. This information can be incredibly useful for general managers to help them predict when players will be drafted, facilitate trades to pick their most desired athletes, and to help them make an order for drafting decisions.

Data set

The dataset consists of information on 3,477 players who were registered for the NFL draft from 2009-2019. The dataset also includes information on players who went undrafted with all of their draft statistics but we choose to remove these players from the dataset and only look at the values of drafted players, leaving 2,254 players. There are 18 categories of data for each which are as follows: Draft Year; Player Name; Age; School; Height; Weight; 40 Yard Sprint Time; Vertical Jump; Bench Press Reps; Broad Jump; Agility 3 Cone; Shuttle; Drafted Team, Round, Pick and Year (combined into one field); BMI; Player Type

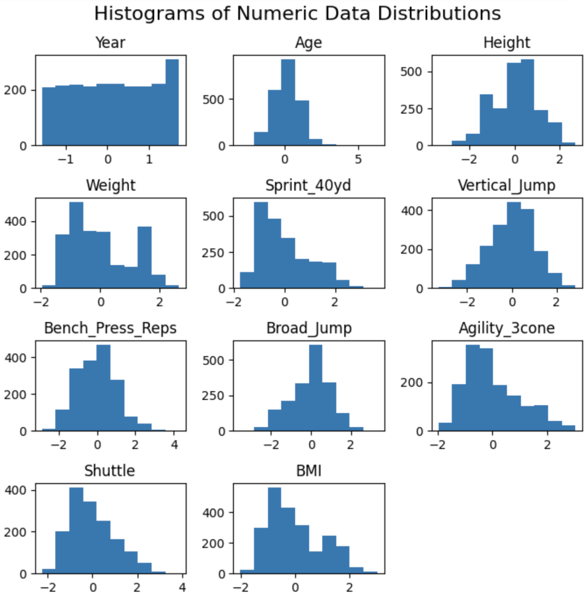

Histograms of Numeric Data Distributions

Figure 1

(offense or defense); Position Type; Position; and Drafted Status. The graphs in Figure 1 display the standardized distributions of the numeric data.

Data Cleaning

The first steps to clean the data included splitting the drafted team, round, pick and year column into four separate columns, removing any stray spaces and converting numeric strings into integer values. This was done to isolate the pick value, which is our goal for prediction. Six of the features contained NA data entries. In the regression, these features were dealt with by creating an additional column that contained a 1 if the corresponding value in the row was NA, and a 0 otherwise. We then set the NA values in the dataset to 0, which allows the regression to take NA values into account without reducing the impact of the other values. The data consists of both numeric and categorical data, which requires use of several different tools during the analysis of the data. The numeric data was standardized using the mean and standard deviation for the feature to ensure that the regression weights could be compared without being influenced by the magnitude of the values. To run regression on the categorical data, we converted the data into numeric values using one hot encoding. For each of the 4 categorical data columns (School, Player Type, Position Type, and Position) a new column was created for all unique data values. These columns were then populated with a 1 if the original column had the corresponding value and a 0 otherwise. This could add up to 7 columns for position type, 3 for player type, 20 for positions and 205 for school for a maximum of 235 one hot encoded columns, resulting in up to 253 columns of usable data for the regression.

Approach

We first decided to look at the full dataset with all of the features to see if we can find the best indicators of draft position for players. Following this analysis, we explored the best features for prediction on a subset of the data such as by a singular position. After cleaning the dataset, we first decided to use ordinary least squares linear regression to inspect the regression weights of the model. This algorithm fits a vector of weights (w) to the data feature matrix (X) to minimize the sum of the squared differences between the predicted value and actual value. The predicted value is obtained by multiplying the matrices X and w. The actual value is the draft pick in which the corresponding player is selected. When analyzing the regression, we looked at the mean square errors for both the test and training data. This allowed us to inspect if the model is overfitting or underfitting the training data as well as give us an idea of the accuracy of the model's predictions. We also looked at the largest absolute value regression weights. These values are important because they have the biggest impact on the predicted value.

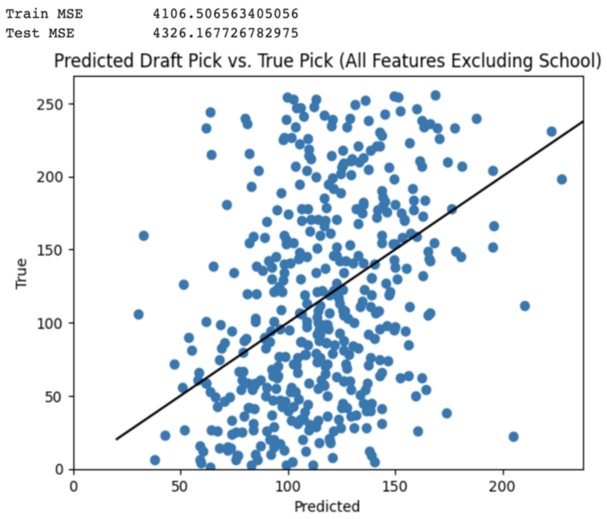
For another measurement of the relative importance of each of the features in the dataset, we calculated the Mean Decrease in Impurity (MDI) of each feature. The MDI goes through every split of the decision tree that uses a certain feature and averages the decrease in impurity when splitting on that feature. The higher the MDI, the more important the variable is in a prediction model. After inspecting the OLS regression weights and observing the results of the MDI experiment, we have a good set of features to suggest to general managers to give a heavier weight when conducting draft analysis.

Analysis

1. Ordinary Least Squares Regression

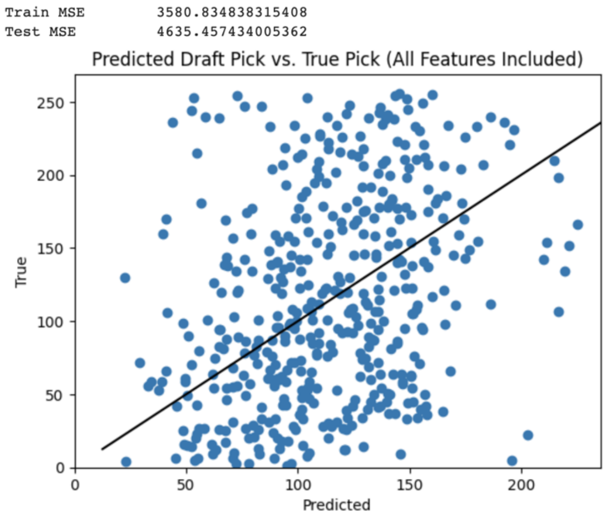
We first ran the OLS regression on all of the features, which gave the results in figure 2. The mean square error is 3580 for the training data and 4,635 for the test data. The large difference in the training and test MSE value indicates the model is overfitting the data. When analyzing the weights for this model, we found the 10 most impactful features were all schools.

Leading us to believe that using school as a feature results in overfitting the data. The largest regression coefficient is 138 and it is for Saginaw Valley State. There is only one player drafted from this school (pick 236) in the dataset and since he was drafted so late in the round we believe the model is overestimating the coefficient for this school due to the small sample size. This small sample size issue is prevalent throughout our dataset, so we decided to remove school from our analysis and we reran the model for figure 3.



```
Train MSE        4106.506563405056
Test MSE         4326.167726782975
```

Figure 2



```
Train MSE        3580.834838315408
Test MSE         4635.457434005362
```

Figure 3

The training MSE was 4,106 and the test MSE was 4,326. Removing school helped to correct some of the over fitting, which demonstrates that school should not be factored in when determining draft order. Additionally, since we have 2,254 athletes in the dataset, having a test set MSE of around 4,326 is a good score for the dataset. This means that on average predictions for when players are going to be drafted is off by around 2 picks on average.

We inspected the largest regression weights to see which features had the most significant impact on draft picks and found that the top 10 features were weight, DB, Offset, BMI, Age_NA, FB, QB, special_teams, Sprint_40yd, and LS. DB, FB, QB, and LS are all positions and indicate that the role you play on a football team despite your combine stats can be a significant factor in your expected draft position. Quarterbacks (QB) and Defensive backs (DB) have negative weights which indicates that they are more likely to be taken in the earlier rounds. These results make intuitive sense since in practice Quarterbacks are often some of the first picks in the NFL draft and special teams players (with a large positive regression weight) are usually late round picks. Of the top 10 regression weights by magnitude, only one of them is a result from the NFL combine (40 Yard Sprint). The presence of only one regression weight indicates that general managers should pay significant attention to a player's 40 yard dash time and that there are many more factors that are important for the NFL draft aside from the combine results. Having only one combine result be a leading motivator for draft position suggests that combine results should be used as a supplement to external analysis. The 3rd most important regression weight was the control term (Offset) which picks up for omitted variable bias which we expect to be significant given the possibility of adding new data points for each college athlete.

2. Mean Decrease Impurity

Following the OLS results, we ran a Mean Decrease Impurity test to have another measurement of the importance of each feature. From inspection of figure 4, we observe that BMI, Weight, Sprint_40yd, Bench_Press_Reps, and Broad_Jump are 5 significant features. We notice here that position is relatively low on the MDI graph which is in contrast to OLS regression weights for positions like Quarterback and Fullback. Consequently, position is only a relevant factor for a player's draft stock if they are in a certain position (such as Quarterback or Kicker). On average, a player's position is not going to have a large effect on their draft stock unless they are in a certain position that would have a large magnitude for the OLS regression weight. The MDI analysis is particularly useful because it informs NFL general managers that the most important factors to inspect from the NFL combine are the 40 yard sprint time, Bench Press Reps, and

Broad Jump length. Additionally, general managers should pay attention to a players weight and BMI since they lead to large changes in the purity of the decision tree.

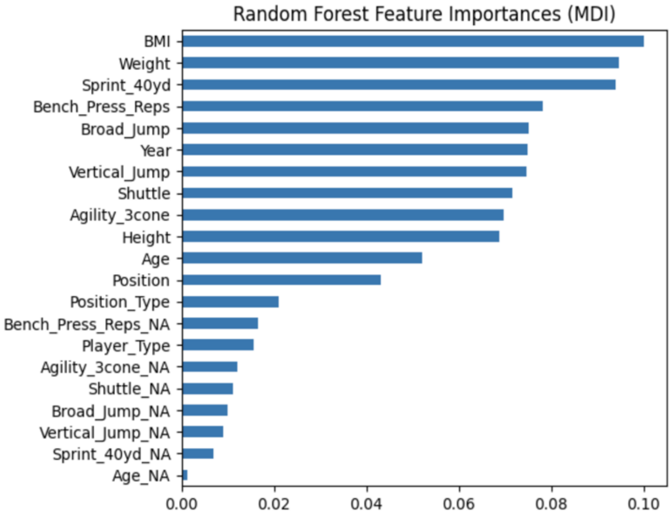
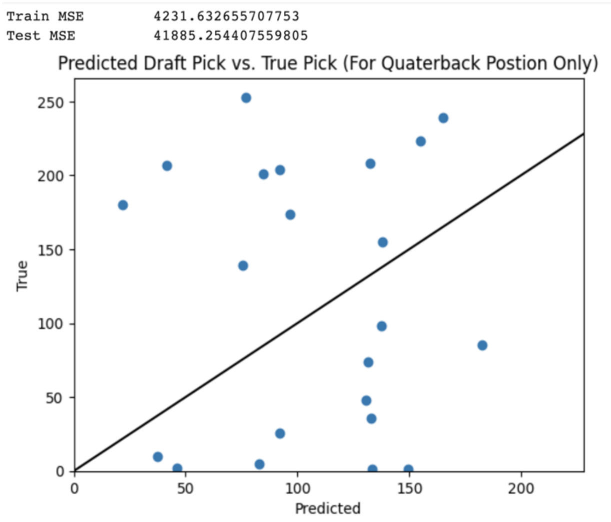


Figure 4

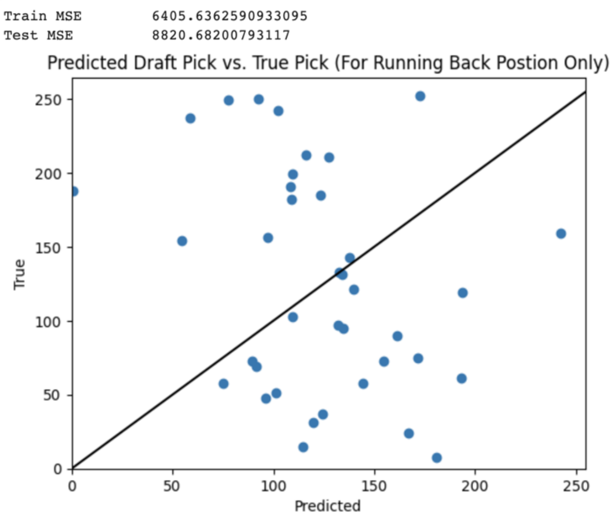
3. Position Based Analysis

The model was unable to accurately predict draft position when the data was filtered by position or position type. The training MSE for the position specific data sets is similar to the training MSE on the whole data set, however, the test MSE was significantly larger (10x). This shows that the model is overfitting the test data which is a result of using too small of a sample size. Figure 5 shows the results from regression on only quarterbacks. There were 5 players in the test data who were drafted in the top 10 picks, however the model predicted their pick to be between 40 and 150. Figure 6 shows the regression on only running backs. While this model has better accuracy than the quarterback regression, the training and test MSE are still larger than the original model. Given the high test MSE values, the model should not be used on individual positions.



```
Train MSE        4231.632655707753
Test MSE         41885.254407559805
```

Figure 5



```
Train MSE        6405.6362590933095
Test MSE         8820.68200793117
```

Figure 6

Weapon of Math Destruction

A Weapon of Math Destruction is defined as a predictive model whose outcomes are not easily measurable and whose results can create a negative cycle that disadvantages people in the dataset. We believe that our project is not a weapon of math destruction because the factors our model uses are objective and do not favor a specific group. Some of the features in the dataset are uncontrollable such as a players height and age, but other features are possible for the players to improve on if they train for it. Although our model uses age, which is a protected feature, we are using empirical data and our model does not disadvantage players based on their age. The model does not tell teams to not draft players based on their age, which would be unfair. This logic applies to all other potential personal characteristics such as height and weight. We believe that our model is fair because it does not disadvantage or discriminate against any group of people.

Confidence in Our Results

When comparing the magnitude of the regression weights with our Mean Decrease Impurity analysis we noticed that Weight, BMI, and 40 yard Sprint time all make significant contributions to a players draft position. This increases confidence in our analysis with the information we have available and bolsters our suggestion for general managers that they should take interest in these factors when constructing their draft strategy. While we would not recommend using our model alone as a predictor of NFL draft order, we are confident that our findings are valid. The factors we found to contribute the most to draft order make sense when looked at with an understanding of football. We are not confident in our results on the position or position type level. When narrowing the data down to players of only one position or position type, there is not enough data which results in a very high MSE.

Conclusion

We would not recommend using this data to make draft day decisions; however, we believe our tool is a starting point for a more accurate model. Looking at our top 10 OLS regression weights only one of them was a test from the NFL combine. While there is a general trend in the data, there are also many outliers which tells us that NFL combine data is not enough to accurately predict draft order. Factors like college stats, number of times a name appears in social media posts, and team needs could also be important factors when determining draft order. With additional resources and access to more data, adding these features could make our tool more reliable.

References

1. https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html
2. https://www.kaggle.com/datasets/redlineracer/nfl-combine-performance-data-2009-2019?select=NFL.csv