Bellevue University

Winter 2020

Predicting NFL Draft Position from Historical Scouting/Draft Data

26 Feb 2021

Kevin Angotti

Executive Summary

The National Football League, or NFL for short, is a multi-billion-dollar industry. The sport has followers worldwide, and the season spans almost the entire year when you dig into each of the different events. One of the largest events outside the Superbowl is the NFL draft. Preparing for the NFL Draft is another event called the NFL scouting combine. This event brings the top 300 or so collegiate players in the nation to Indianapolis, Indiana, for what is essentially a 4-day job interview.

Hopeful players compete in different drills and formal interviews for team scouts and General Managers or GMs for short. These drills and interviews can impact where a player might be drafted in the Draft. Teams have to analyze hundreds of players in a few short days to determine if they are worthy of a team's select draft picks. Every year in early March, the event can be seen as the last chance players can be evaluated by teams before the Draft.

Looking into different aspects of player evaluations might explain where a player could be drafted. Many player rankings usually revolve around college statistics. However, will it be possible to take that historical player ranking data and determine if the player's ranking can be improved based on similar rankings over time?

Preparation

This project looks at various predictive analytic methods to understand relationships within the data to identify factors that could help improve or reduce a player's overall value, thus affecting where said player might be drafted. Key factors examined during this project were each of the NFL combine drill statistics and player measurements. The main key factors were height, weight, bench press, vertical jump, broad jump, shuttle, cone, and 40-yard dash. Secondary key factors were the player's AV, round, and pick.

Initial Findings

During this project, initial findings yielded several interesting returns; more on this regarding findings can be found below. Each of the hypotheses guiding this project were put through the modeling process. Upon completing the modeling process, the outcome points toward predicting a change in draft pick and player value based on the input variables. This allows for the analysis of future drafts by using up-to-date player statistics.

Abstract

The Nation Football League is one of the most popular professional sports in the world. Every year from March to February of the following year, the sport takes center stage in television market ratings, sports betting, stadium ticket sales, merchandise sales, and much more. Outside of the Superbowl, no other NFL event is more televised than the NFL Draft. The NFL Draft is a vital part of opening the season. However, one other event could be seen as just as essential and draws attention to the sport's media and fans, the NFL Scouting Combine. This event occurs in early March and is held for 4 straight days to evaluate the player for that year's Draft [12].

The NFL Scouting Combine is, in short, a job interview for the sum of 300 plus players coming out of college to try and be drafted into the NFL. Players are pre-ranked based on their collegiate playing statistics and measurements. The combine can improve or reduce that ranking. In years past, players have opted out of the combine for fears of injury or reducing their projected rankings. However, this usually does not bold well for teams looking to evaluate players.

Teams have access to player stats at the touch of a button. However, not many analytical features allow a team to compare players to historical data for determining actual value. There have been many players throughout history that have either been drafted extremely high or extremely low. One of the most notable players is Tom Brady drafted 199th overall in the 6th round. If you were to analyze his college stats and measurables today and compare that to similar players, would his pick location change? Looking into how a player's current actual value (AV) stack up to players of the past and where they were drafted is one area that could be analyzed.

This project aims to determine the best outcome a player will have compared to historical player statistics and what factors might help determine draft pick location and actual AV value. Another hopeful outcome is determining if a tool can be created from predictive analytics for more straightforward player analysis.

Keywords: Predictive Analysis, NFL Scouting Combine, Linear Regression, Decision Tree, EDA

Introduction

The National Football League is an American professional sport that has been around since the early 1920s. Considered a business that draws billions of revenue a year from different major events allows for many different analytical approaches. This project focuses on looking at players who are just entering the league, hoping to be drafted by one of 32 teams. Each NFL team has a 52-man roster with an additional 10 man practice squad allowing each team to hold

62 players. The NFL draft brings in late April, and each year roughly 260 new players enter the world, that is the NFL.

Players are evaluated throughout their collegiate career and, once again that the NFL scouting combine. A player can improve their draft value by being roughly 300 players invited to participate in the 4-day scouting event. The NFL scouting combine puts players through different drill sets based on the position the player is in. However, all players participate in five main drills before heading to their primary position drills. The 40-yard dash, vertical jump, shuttle cone, and bench press are all drills every player completes as their main measurables.

Players can be valued pre-combine high and still provide a good showing at the scouting combine only to not live up to their ranking in the NFL, while others that may have had lower values pre-combine become star players in the NFL. How can players truly be evaluated based on current stats and measurements? There is a need for players to be evaluated further. This project looks to do just that and take player analysis currently used and subject that to historical data to predict real value.

Background Information

Methods

Data Understanding

When approaching the data for this project, I began by examining variables within the dataset. This takes time due to the dataset length of both variables and rows; I used a combination of R programming language, Python programming language, and Power BI. To begin the project, I turned to R; this language is an excellent tool to quickly look into a dataset to conduct summary statistics and see initial trends and what each variable might have to tell say.

A quick linear regression model was used to see what these trends might mean after conducting the summary statistics. The R programing language provides a simple plug-and-play tool for different LR models. This is useful to get an idea of what your variables might have to tell you in a more invasive model setting. I took the initial LR model from R and my summary statistics. I used that to create my initial visuals in Power BI. Doing this at the beginning of the project allowed me to see what the data looks like before running any Python modeling.

Initial visualizations from Power BI provided evidence that the AV or Pick variables may be a possible target for the modeling. Other areas of interest were to review each of the measurable metrics and analyze that against historical data. I then used Python to conduct all of the EDA and create the models to run the data against. After the EDA was completed, the first model I created was linear regression which used the Ordinary Least Squares or OLS for short. OLS looks at the relationship between the dependent variable and at least one of the independent variables chosen. I then followed this up with a decision tree to further correlate OLS variables to be valuable for prediction. These models provide insight into the research questions, allowing for future analysis and a potential application or tool that teams can use to evaluate better players entering the Draft.

Data Preparation

Extensive data preparation was performed on the dataset to transform key categorical predictors into a numeric form. This was done to prepare the data for modeling and further analysis through EDA. Basic visuals were completed beforehand using Power BI to determine which variables could be used for modeling from summary statistics. During the data cleaning process, certain variables were transformed from objects to numeric values. This would allow for the summary statistics, linear models, and decision tree to run and provide insights into the data.

The variables were transformed in Python on the data set by the use of the formula, (df['columnName'] = le.fit_transform(df.columnName.values). A new dataset created with the transformed variables was saved and then loaded into the Rstudio to run the summary stats and the initial liner model. The main variables that were transformed were Forty, VerticalJump, BenchPress, Cone, Shuttle, Round, Pick, Year, Height, Weight, And AV.

The next process I conducted was to create each of the highlighted variables visuals to get a better look at the data. As you will see in Appendix A, Figures provide a good sense of what the data is currently suggesting. These visuals were used to compare the data before it was transformed for modeling. Each of these variables was not normalized for plotting purposes.

Modeling

The extensive data preparation allowed for a linear regression model to be performed. R's initially built linear model was formulated with the Pick variable as a function of Height, Weight, VerticalJump, Cone, BenchPress, BroadJump, and AV variables. These variables returned useful insights for significance, and all but three of the variables returned a p-value under the 0.05 threshold. The overall P-value of the model yielded < 2.2e-16 or (0.00000000000000022) with an F-statistic of 117.8 and an adjusted R-squared of 0.31.

I then shifted to Python to run the primary linear model based on the summary statistics and initial LR model. I utilized OLS regression to understand better the variables I was working with for prediction and get a better result out of my predictions other than a typical linear model.

Finally, I run the data through a decision tree algorithm based on the OLS regression model results to further analyze the prediction of the key variables found during the analysis.

Running multiple model types provides a good range into what the data could tell us, thus allowing for sound analysis.

Results

From the preliminary results, things are looking promising. While further analysis is needed to draw definitive conclusions, R's initial model results are shown below.

Linear Model

```
lm(formula = Pick ~ Ht + Wt + Vertical + Cone + BenchReps + BroadJump +
    AV, data = combine_data_since_2000_PROCESSED_2018_04_26)

Residuals:
     Min       1Q   Median       3Q      Max
-131.318  -44.327   -5.144   41.507  227.227

Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 357.2872359  74.0866541    4.823  1.53e-06 ***
Ht           -0.0007102   0.8464353   -0.001   0.99933
Wt           -0.3842575   0.0916117   -4.194  2.86e-05 ***
Vertical     -0.1558617   0.5656389   -0.276   0.78292
Cone         16.3364993   6.1006804    2.678   0.00748 **
BenchReps    -0.4202100   0.2902075   -1.448   0.14780
BroadJump    -1.8682955   0.3005428   -6.216  6.27e-10 ***
AV           -4.1953720   0.1685477  -24.891   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.92 on 1854 degrees of freedom
  (4356 observations deleted due to missingness)
Multiple R-squared:  0.3078,  Adjusted R-squared:  0.3052
F-statistic: 117.8 on 7 and 1854 DF,  p-value: < 2.2e-16
```

Table 1. Linear Regression Model

After running this regression model, which used different summary statistics to determine significance, such as an F-Statistic, P-Value, and Multiple R-Squared, these measures only provided some of the information needed to determine if the data could predict on a model. I also created visuals in Power BI (Appendix A) from these statistics to get a peek at what a model might conclude before I started building my models.

Shifting to Python, I started with some exploratory data analysis (EDA) to get insights into the key variables and run some other linear regression and random forest models on the data. Appendix A will show some of these results and be further discussed in this paper's discussion section. After the EDA was completed, I started to code the main models for the analysis. I first created an OLS regression model that follows the linear format for predicting the target. However, the main difference here is it looks at the dependent variable and compares it to one or more independent variables. This type of regress was conducted twice or in two separate models. The first ran the entire measurable statistics height, weight, cone, vertical, broad, bench, shuttle, forty, and some draft data round, pick, and AV. The results of this model returned the variables that would be best used for prediction. The variables used in the second OLS model were height, forty, broad, shuttle, cone, and AV. From the second model's results, I was able to predict if a player's pick or even value would increase or decrease. The prediction part of the model can be adjusted to look at a single-player or position group.

Suppose a player like Tom Brady thought of one of the most winningest quarterbacks of all time, who was picked late in the NFL Draft round 6 pick 199. This prediction model would have him gaining 34.7 spots in the 2000 NFL Draft to pick 164.

The final model created for this project was also built using Python was a decision tree. This model took in the final OLS regression model variables to create the tree. A few more conversions of data was needed to allow the tree model to run. The changes made transformed the larger floating decimal values to whole numbers rounding up their values to the nearest number. The tree looked at the Pick a player would be based on the predicted AV change. The resulting tree showed an accuracy of 30 to 35 percent, with the lowest run at 25. The tree takes in the root node of Pick less than or equal to 133. Depending on where the team or scout would like to value a player would determine the choice here. If looking for value in later rounds, i.e., higher than pick 133, you say no and move to the right of the tree to the next node. This next node is less than or equal to Pick 191. In this case, you would look at the leaf on the left of the right side and get a Gini score of 85 percent and have players to use for evaluation. Though using a dataset that has players from 1987 to 2015, this decision tree could easily be modified to bring in either more recent draft years or filtered for just the last 10 years. This could become a future step in the project.

Looking forward to the next steps, I would like to obtain a software engineer's help to adapt these models into a phone application that NFL teams could use when attending the NFL scouting combine.

Discussion

During this project, insights gained from summary statistics allowed for initial analysis to be conducted. This is done to determine which variables could be used for prediction. This project started out with a single dataset that contained 16 variables and with over 6200 rows of data. After running through each of the project's steps all the way through modeling, the results though promising did not feel complete. I set out to find more data that could be used to help tell the full story.

The initial dataset came from the Kaggle website and provided an excellent start to the questions I was looking to answer. I found two more datasets that would later replace the initial dataset from DataWorld. The new datasets provided separated information for the NFL combine statistics and NFL draft Statistics. The combined dataset had 16 variables and 10 thousand rows, while the Draft dataset had 33 variables and 8K rows. Upon combining the two, and the resulted dataset used to finish this project contained 47 variables and 5600 rows. More cleaning and EDA needed to be conducted. However, this was not too much additional work as the foundation had been created from the initial dataset.

The newly merged dataset allows for better accuracy during the modeling stage. The EDA process on the final dataset returned excellent insights to the questions I was looking to answer. The 40-yard dash is one of the major event drills at the NFL Scouting Combine, and from the EDA, we can see that height and AV have a positive trend associated with one another. You can see in figure 1 below the correlation between the two variables.

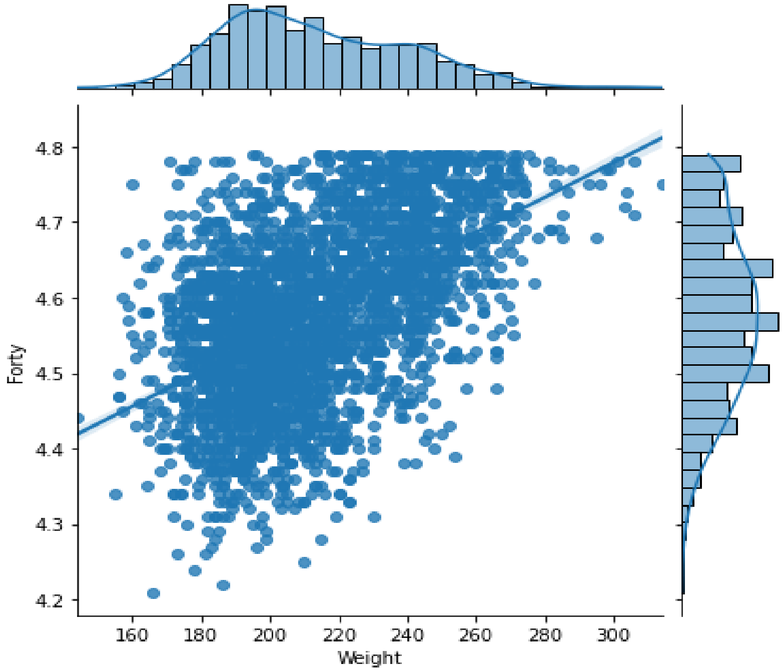


Figure 1 40 Time vs. Weight

The ideal weight and 40 times seem to be around the 200 pounds and 4.5 range. The correlation matrix shown below shows the main variables that have significance in a player's measurables at the draft.

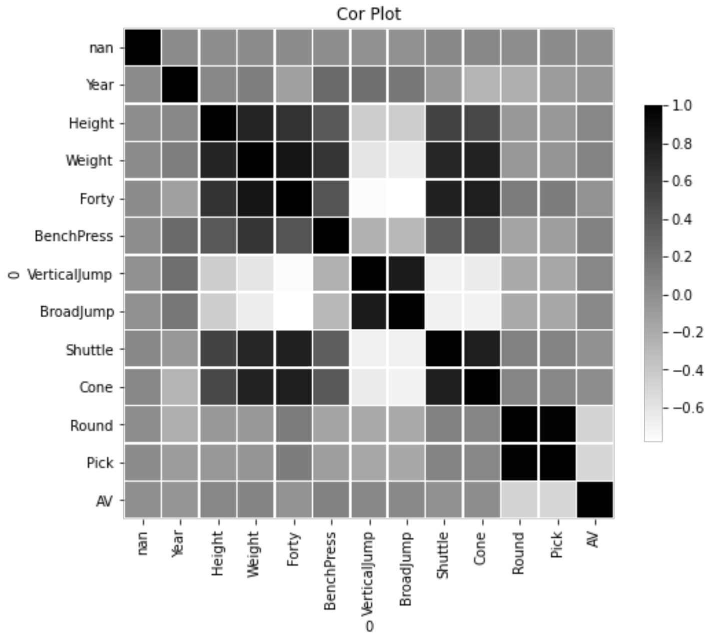
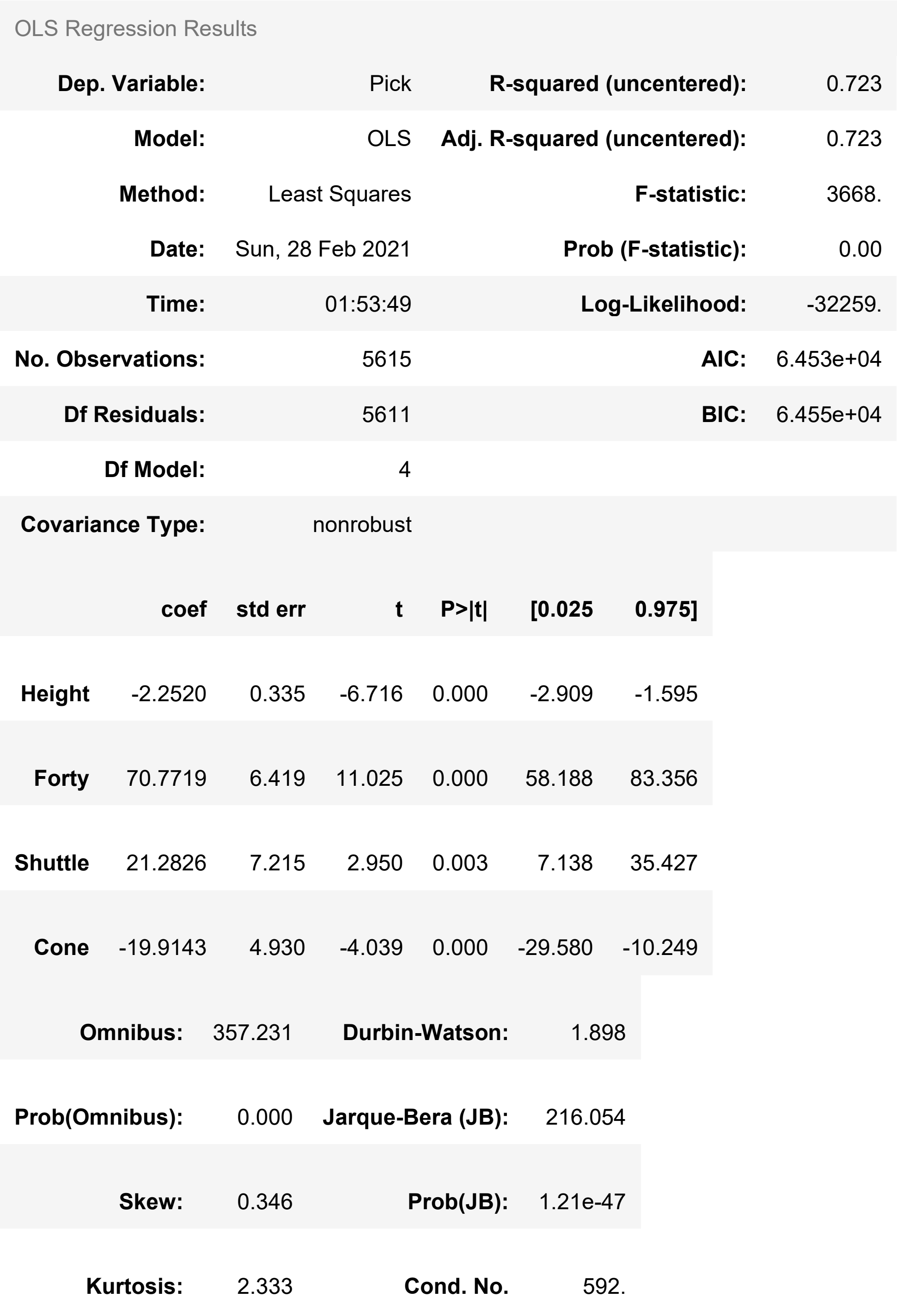


Figure 2 Cor Matrix

Here, you can see that height, weight, forty, shuttle, and cone are the variables that I would use in my modeling. Most drafted positions are highlighted b what's known as the offensive skill positions from the EDA process; I was able to look at these positions against variables like height and weight. Figures 4 and 5 Appendix A shows that ideal measurements in the two variables for the QB, TW, RB, FB, and WR positions are 72 to 77 inches tall and 225 to 250 pounds.

The OLS regression model I ran looked at each of the variables found during the summary statistics process. Two were created for this project, with the first looking at all of the variables, which allowed me to narrow down the list to just the following four variables for prediction: height, forty, shuttle, and cone, four of the five found in the correlation matrix. The results of the final OLS regression model can be found in figure 3 below. Though the R-squared dropped from 98% to 73%, the condition of the model improved significantly.



OLS Regression Results

Dep. Variable:	Pick	**R-squared (uncentered):**	0.723
Model:	OLS	**Adj. R-squared (uncentered):**	0.723
Method:	Least Squares	**F-statistic:**	3668.
Date:	Sun, 28 Feb 2021	**Prob (F-statistic):**	0.00
Time:	01:53:49	**Log-Likelihood:**	-32259.
No. Observations:	5615	**AIC:**	6.453e+04
Df Residuals:	5611	**BIC:**	6.455e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Height	-2.2520	0.335	-6.716	0.000	-2.909	-1.595
Forty	70.7719	6.419	11.025	0.000	58.188	83.356
Shuttle	21.2826	7.215	2.950	0.003	7.138	35.427
Cone	-19.9143	4.930	-4.039	0.000	-29.580	-10.249

Omnibus:	357.231	**Durbin-Watson:**	1.898
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	216.054
Skew:	0.346	**Prob(JB):**	1.21e-47
Kurtosis:	2.333	**Cond. No.**	592.

Table 2 OLS Regression Model

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Using the OLS model results, I created a prediction that could be used on a single player, a position group, or even an entire dataset. As mentioned earlier, Tom Brady was drafted in the year 2000 at pick 199. Based on modeling, he would have been projected the 162nd pick in the draft, which is the mid-fifth round, versus the sixth round that he was drafted in. The projected pick location is not always an improvement. Players across the board move up and down based on the variables used for prediction. Figure 3 below shows the extensive movement of Picks for every player drafted from 1987 to 2015.

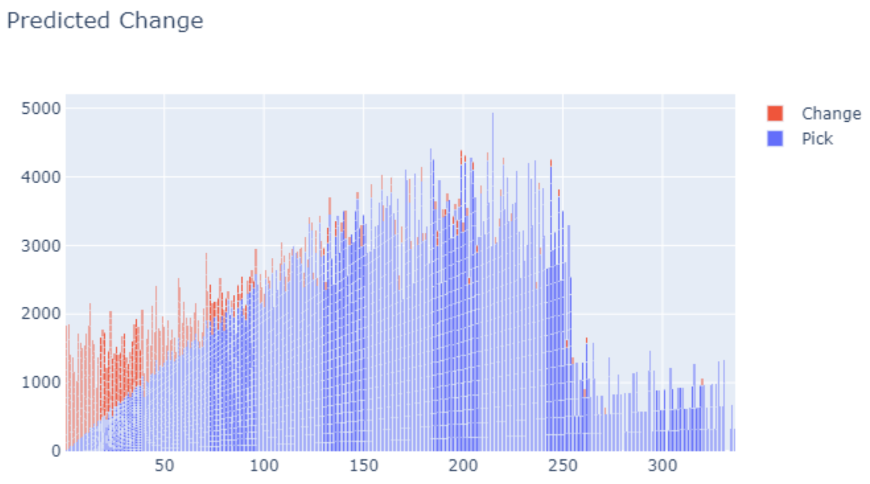


Figure 3 Predicted Change

Additionally, I ran a Random Forest algorithm against the dataset, using the sklearn format. The returned model accuracy was 71%, providing me with more understanding that the target variable was achievable.

The project wrapped up with a decision tree that too provided good scoring. The Gini for the target variable never fell below 72% in any direction the node took the user. See Appendix A for the decision tree results. Appendix A provides all of the modeling visuals and also includes visuals created in Power BI.

Conclusion

To look into NFL historical data and trying to find insights has become a possibility. The analysis I was able to perform is just a small picture of the whole aspect of this project's goal. I want to dig a little deeper and use more additional drafts and combine data from 2016 to 2020 to see if the modeling holds up. From there, reach out to a developer or software engineer to help build the application. A tool of this magnitude could become essential for NFL clubs to use during their analysis of a draft class.

Determining the increase or decrease in a player's draft position was achievable, thus allowing for this project to move forward. Locating or creating the additional datasets that could be added to this one for further analysis becomes the first step into testing the modeling strength. The application could be used similar to a google search where the model results show the players projected draft position based on their performance in college and at the combine. The Players already come to the combine with a grading score. However, this project takes those scores, uses them against the historical draft, and combine data for better predictions. In all, this could be a useful tool that has not been created.

Acknowledgments

I want to thank Professor Catie Williams for her continued guidance and support throughout the project. I would also like to thank Data World for making the project data publicly available.

Questions

1. How much will analysis like this cost for any potential project my company may have?

2. Were there any part of your research questions you were unable to answer?

3. Will there be copyright issues with the plugin application?

4. What type of funding do you have for creating this application?

5. do you plan on taking your project to Kickstarter or even attempt to get on Shark Tank?

6. If you could change anything about this project, what would that be?

7. What was the most significant pitfall of your project?

8. Of the models you used, is there one you liked other than the others?

9. Where there any other variables you considered using as a target besides the Pick?

10. Will the models and other data analysis be available to the public?

References

1. Savvas, T. (2018). Kaggle. Retrieved from https://www.kaggle.com/savvastj/nfl-combine-data

2. Kelly, D. (2018). How To View Grit. The Ringer. Retrieved from https://www.theringer.com/2018/4/12/17227604/nfl-draft-intangibles-scouting-evaluation

3. Woo, M. (2019). How Much Does The Combine Reveal About Future NFL Player? Inside Science. Retrieved from https://www.insidescience.org/news/how-much-does-combine-reveal-about-future-nfl-players

4. Dunlap, A. (2013). Things NFL Scouts Look For At The Combine That Fans Don't. B/R. Retrieved from https://bleacherreport.com/articles/1535607-things-nfl-scouts-look-for-at-the-combine-that-fans-dont

5. NFL Draft Combine Testing. Retrieved from https://www.topendsports.com/sport/gridiron/nfl-draft.htm

6. Witt, T. (2020). Is it possible to predict the success of NFL Draft picks? FanNatin. Retrieved from https://www.si.com/nfl/chiefs/gm-report/predicting-nfl-draft-pick-success

7. King, J. (2020). Using Machine Learning to Predict Fantasy Football Points. Towards Data Science. Retrieved from https://towardsdatascience.com/using-machine-learning-to-predict-fantasy-football-points-72f77cb0678a

8. Fridson, M. (2017). NFL Draft Analysis: 30 years of Player Outcomes. NYC Data Science Academy. Retrieved from https://nycdatascience.com/blog/student-works/nfl-draft-30-years-outcome-analysis/

9. Moore, K. (2017). Predicting NFL Success with Algorithms. Cognitive. Retrieved from https://www.cognitivetimes.com/2017/08/predicting-nfl-success-with-algorithms

10. Robbins, L. (2016). A Data Scientist Dissects the 2016 NFL Draft. WSJ. Retrieved from https://www.wsj.com/articles/a-data-scientist-dissects-the-2016-nfl-draft-1461793878

11. Bronshtein, A. (2017). Simple and Multiple Linear Regression in Python. Towards Data Science. Medium. Retrieved from https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9

12. Scouting Combine. (2021). NFL Scouting Combine. NFL Network. Retrieved from https://www.nfl.com/network/events/nfl-combine

Appendix A

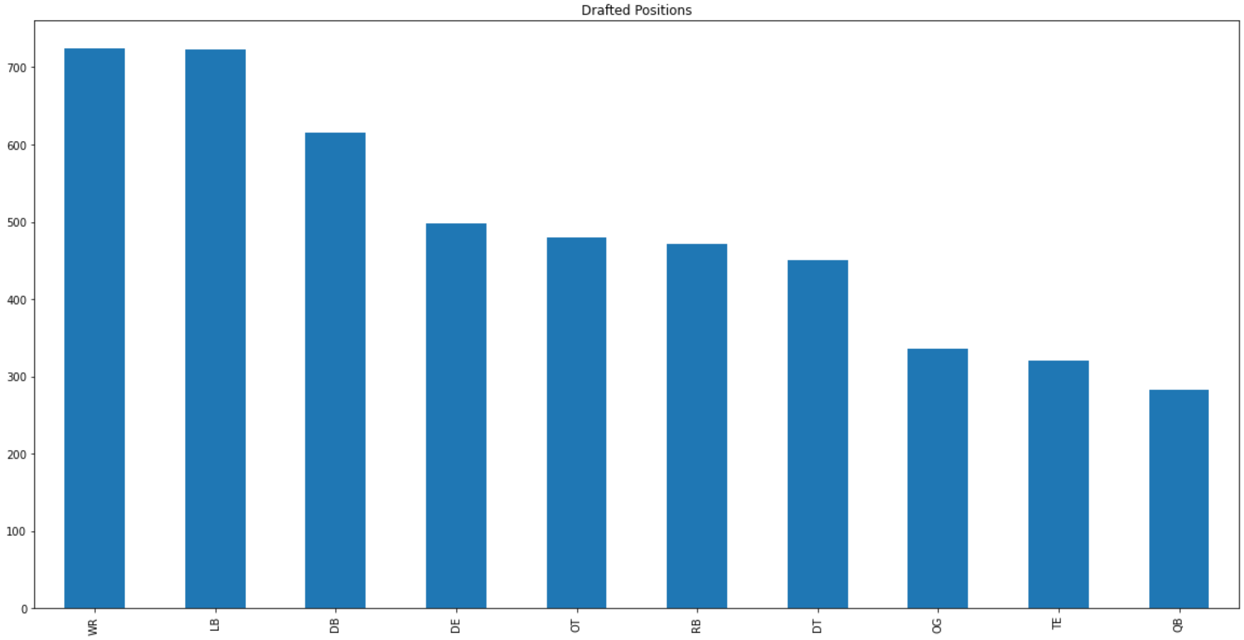


Figure 5: Positions Drafted

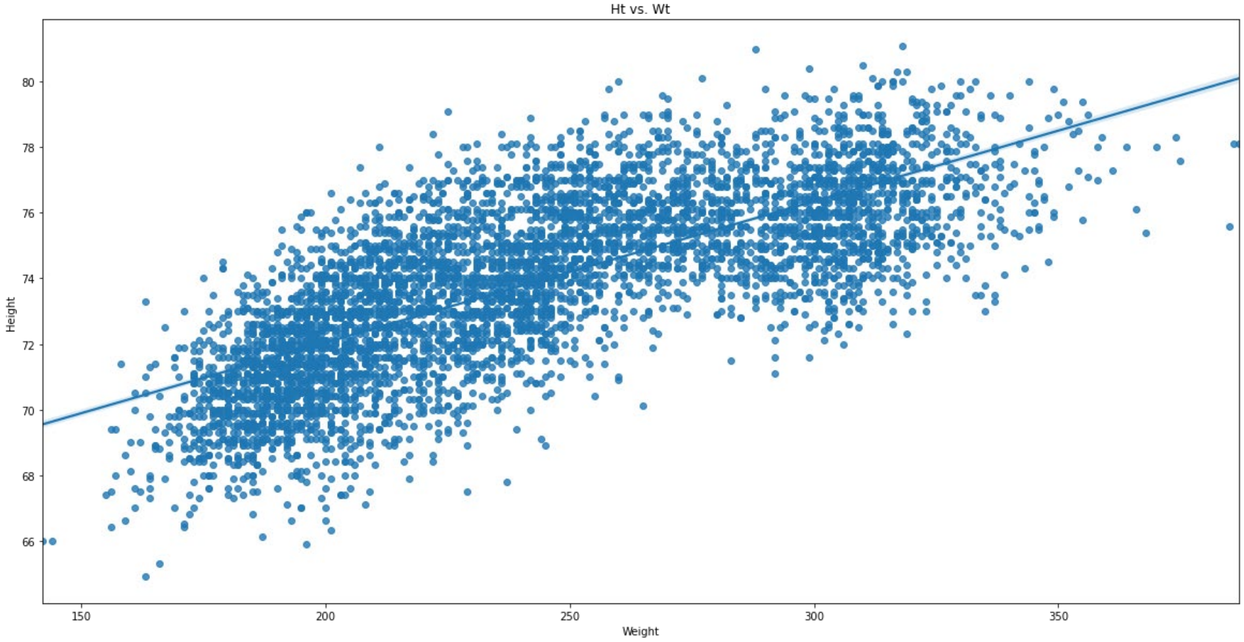


Figure 6: Height vs. Weight

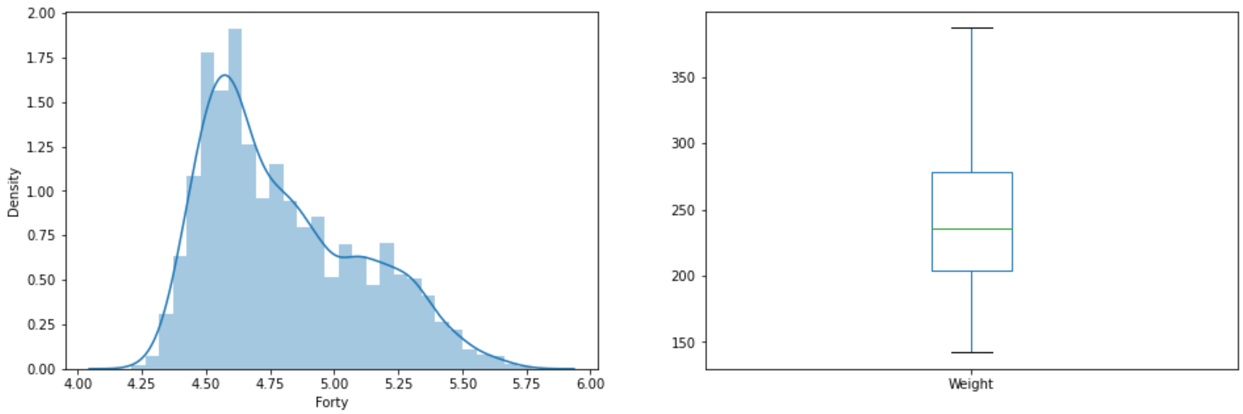


Figure 7: 40 vs. Weight

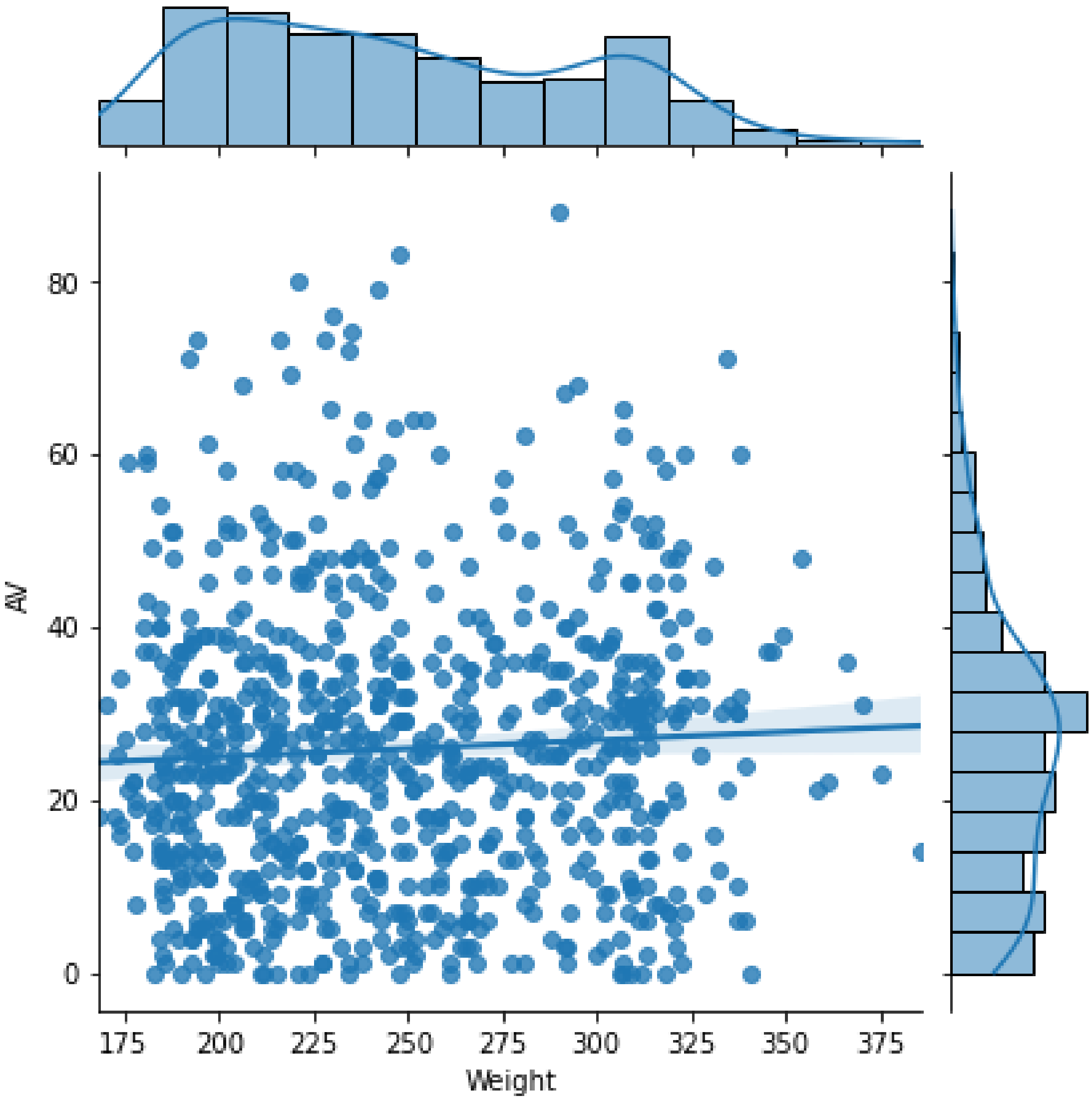


Figure 8: AV vs. Weight

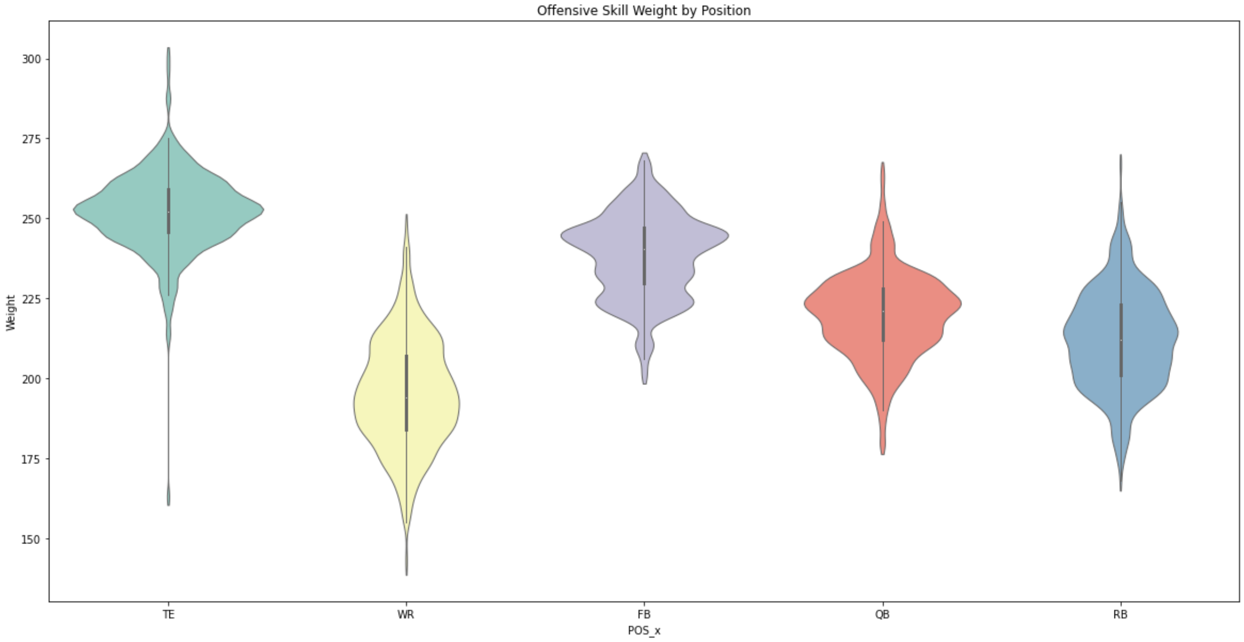


Figure 9: Offensive Skill vs. Weight

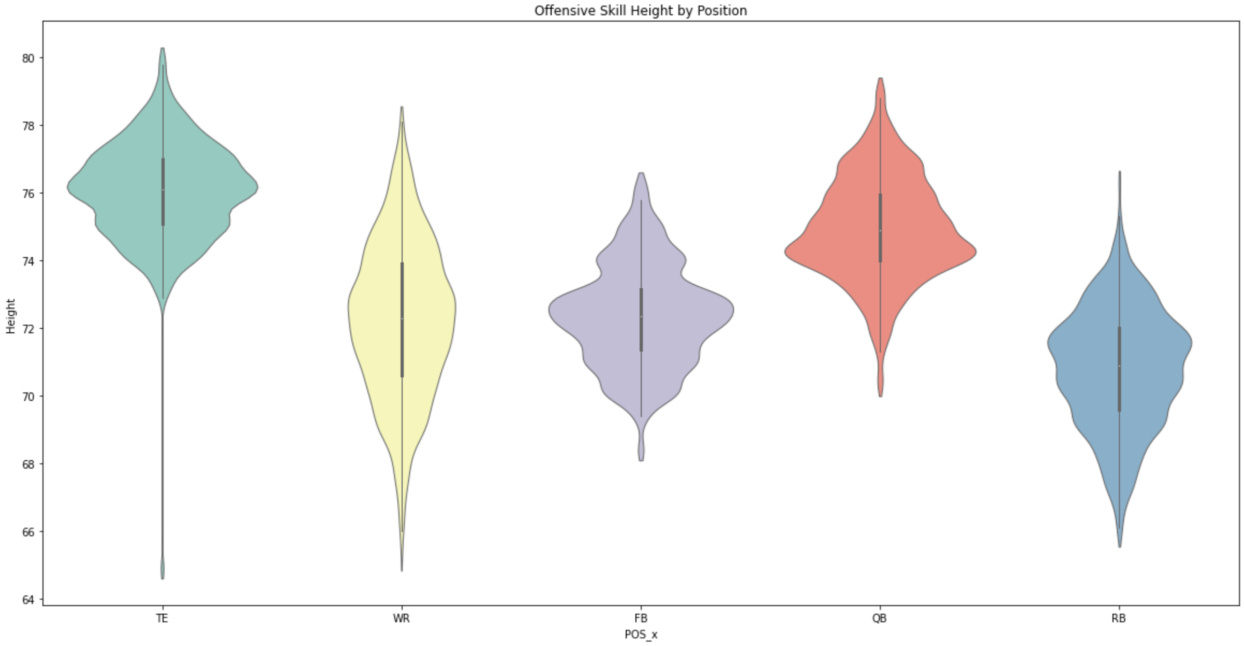


Figure 10: Offensive Skill vs. Height

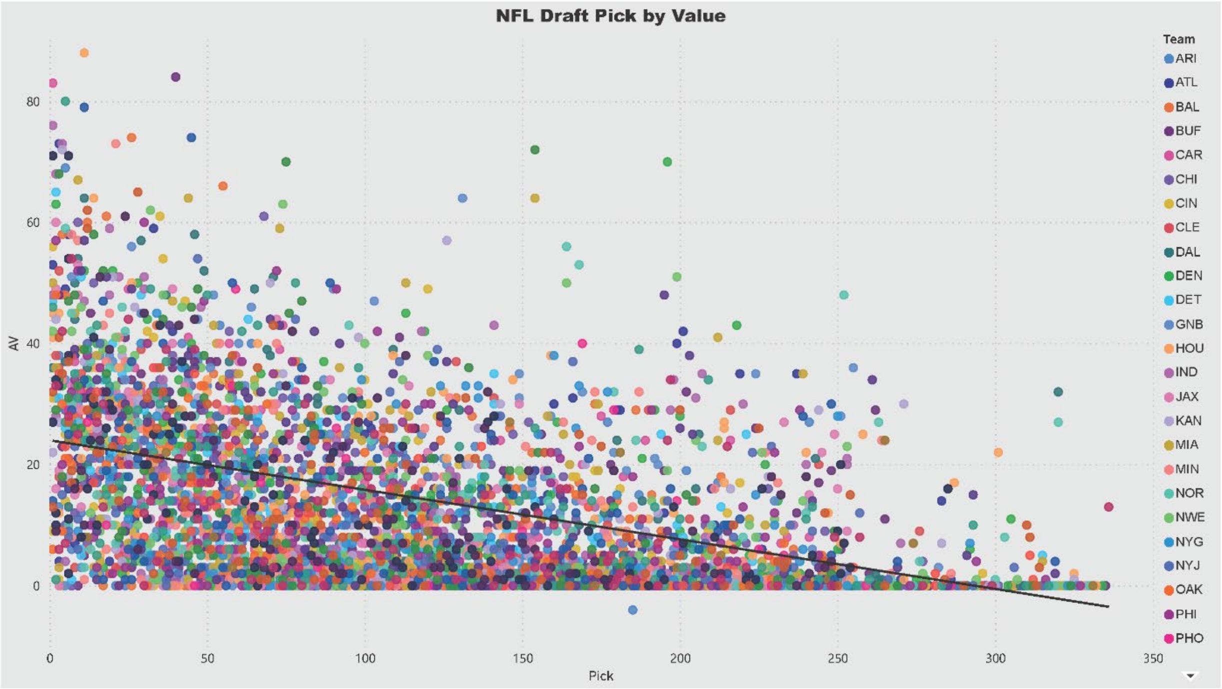


Figure 11: Picks by Value

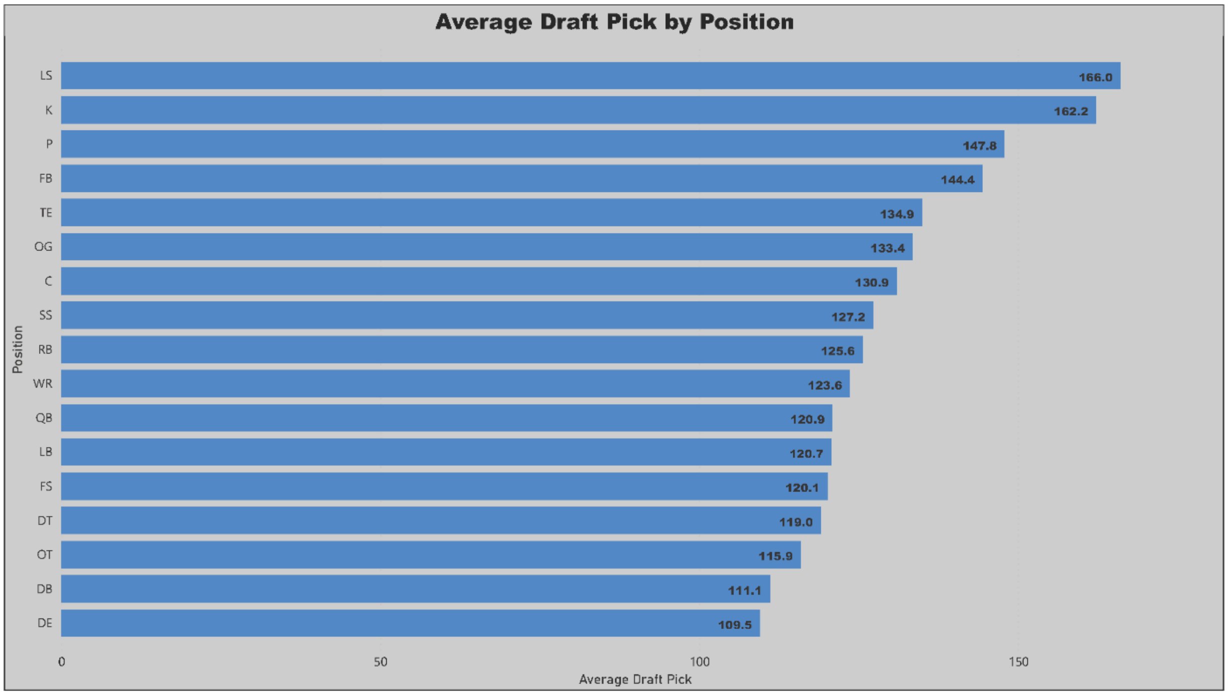


Figure 12: Average Draft Pick by Position

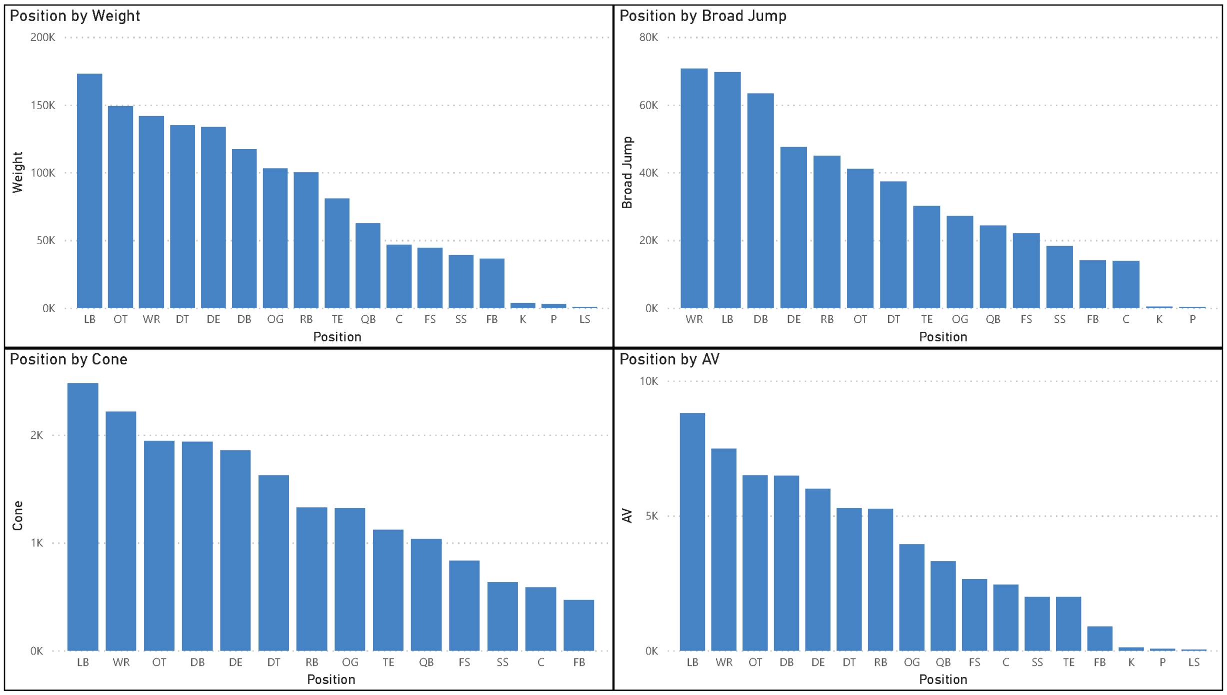


Figure 13 Highlighted Measurebales

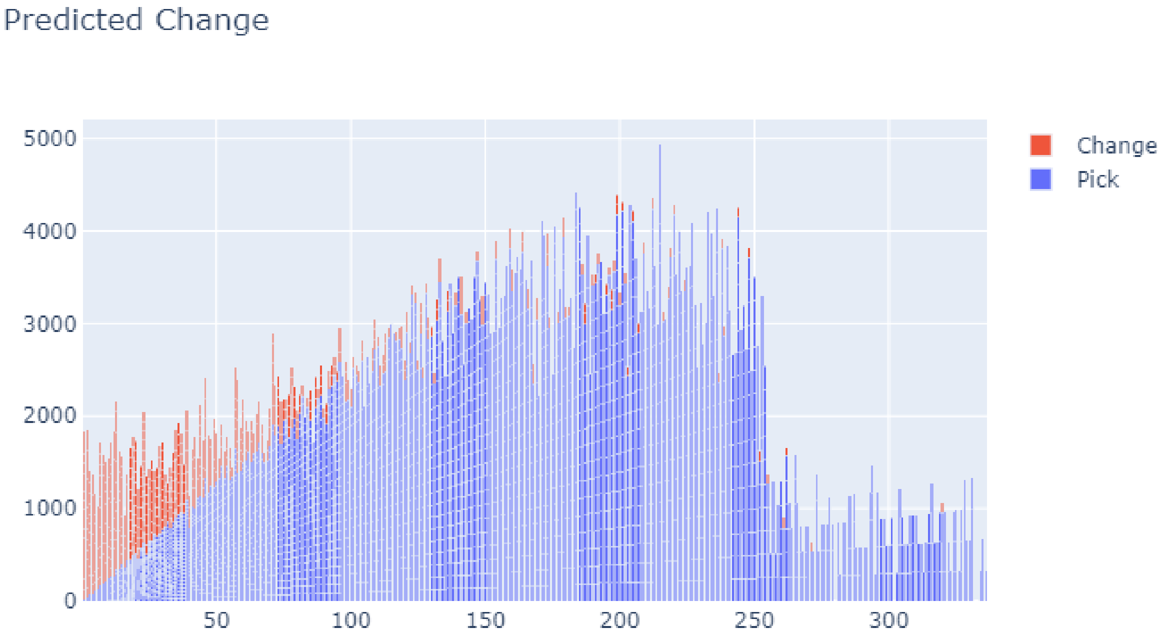


Figure 13 Predicted Draft Position Change

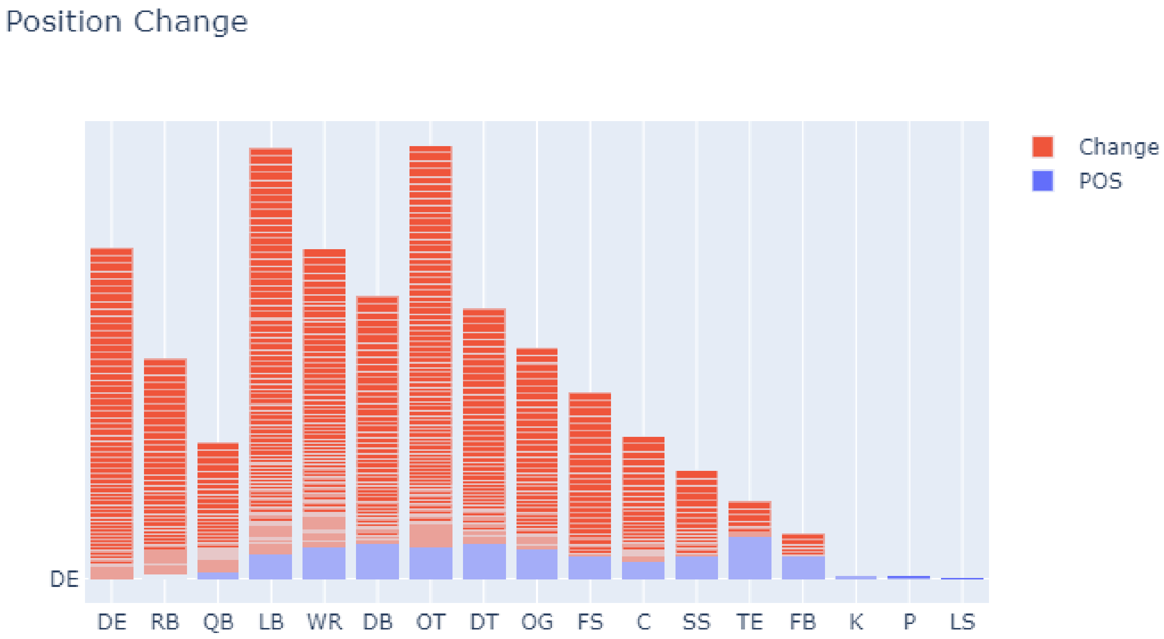


Figure 15 Predicted Change by Position