I. Introduction

This paper explores the relationship between NFL Combine performance and long-term success in the NFL, with success measured by career snap count. The NFL Combine is a 4-day event in which prospective college players compete in a variety of physical tests such as the 40-yard dash, bench press, vertical jump, and shuttle run. Along with a player's college film and private interviews, NFL scouts and organizations rely on these results to make multi-million dollar decisions when drafting players. While these tests are heavily promoted and used by NFL teams and media, there is an ongoing debate as to how well combine results accurately predict an athlete's future success in the league.

In order to study this, we used data from 2016-2023 NFL Combines and compared it with each players' career snap counts; which was our metric for measuring a player's success. Our objective was to find if specific Combine events are more predictive of success along with looking for disparities among different positions. The goal of the study is to find valuable insights into how well NFL Combine results predict professional success and to help inform scouting and drafting decisions from NFL organizations.

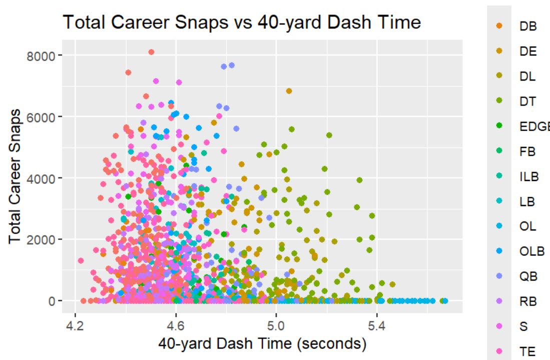
II. Exploratory Data Analysis

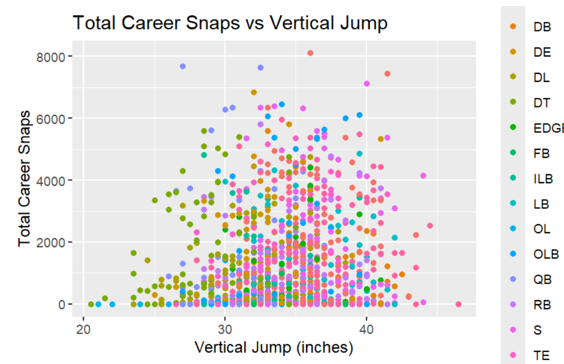
Our first data set we acquired was from Kaggle. It included player name, position, school, and all combine scores going back to 2000. We brought in the other half of our data from Fantasypros. Fantasypros is a leading website in fantasy football data. They have extensive ratings and past data that is a commonly used resource for fantasy drafts. We specifically grabbed player name, position, team, and total snap count. We did this by using scraping techniques for the individual tables, looping the process by year to compile everything we needed into a data frame with all years. We then pivoted by player and year so that each individual only had a single row of data.

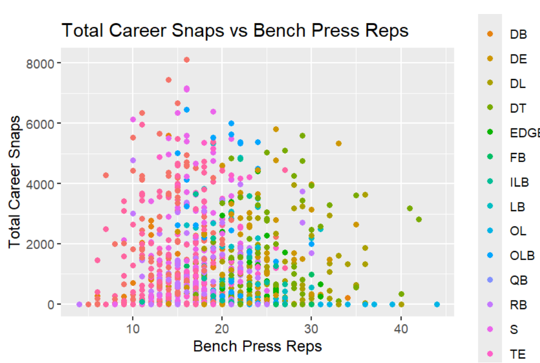
When combining our data sets, we left joined upon player names. We chose a left join because we wanted to make sure that we included all players who went to the combine, not just those who played a snap in the NFL. There is a lot of value in learning what results contribute to just getting snaps on a roster.

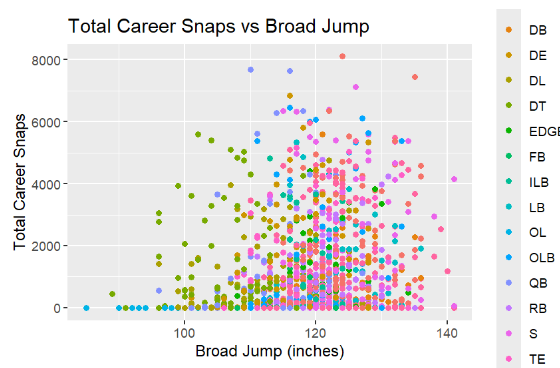
Since there was no ID number in the datasets, it made the join more difficult. There were some name mismatches due to different spellings between the two databases. To alleviate these errors, we made some changes to the names in each table. We capitalized every letter in the player's name, some mismatches occurred because of capitalizations. We also removed all Jr, Sr, II, III, IV, and V suffixes from the names. This removed other forms of mismatching. And finally, we eliminated periods from player's names.

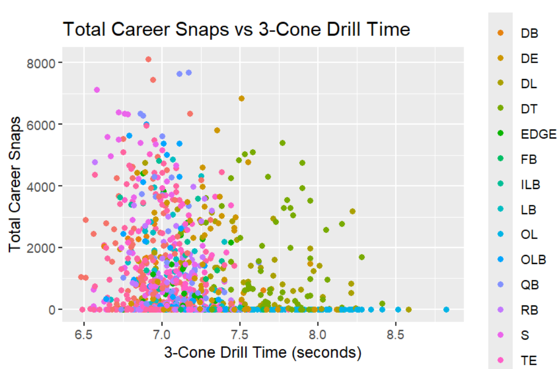
We had quite a few NA's in our data. This was due to the fact that the Combine events are optional. The players can choose to take part in as many or as few events as they choose. To combat this issue we found the mean for the group and went one standard deviation above or below the mean depending on the event. If the event has a desired result that is lower than the mean (ex. 40 Yard Dash), we went one SD above the mean. If the event called for a higher desired number (ex. Bench Press), we went on SD below the mean. We did this because if we were to eliminate rows with NA's we would not have a large enough data set. We did not simply go with the mean because we wanted to punish players who did not participate. A lot of the time players pass on an event because they feel they are not strong in it. We do not want to give them an average score when eliminating the NA’s. The data had the following distributions for example:











To run our Random Forest we made our target variable, snap count, a binary variable. This would allow us to run a confusion matrix for each of the position groups. To make it a binary variable, any total snaps over 50 we gave a value of 1 and anything 50 or under a value of 0. The 1 was our positive value.

For the XGBoost model we used each of the combine events and ran the model for each position. While there were some NA's, a variety of players from each position participated in each event. Using each of the events enabled us to perform a SHAP analysis to assess which events contributed most heavily to the model, and therefore had the most impact on a player's later success.

III. Leaning Algorithm Training and Testing

For all of our models we ignored position, name, and school. School could be grouped by conference and included in a future project, but we wanted to focus exclusively on the Combine measures. One method used to pick the most important variables in our regression and Random Forest models was to view what percentage of athletes performed the test itself. Each position has different rates of participation. We felt that if over 50% of a position group did not participate in an event, we would scrap it as a variable. This led to the elimination of the 3 cone drill for DBs and the 3 cone and shuttle for the RBs.

To start our project we ran a Random Forest model. We simplified the predicted variable and made it binary. The focus of the Random Forest was to see if we could predict if a player would play 50 or more snaps in the NFL. With these predictions, we ran a confusion matrix for all of the position groups.

All of our positions had relatively similar accuracy. Linebackers had the highest with 73%, while running backs had the lowest at 65%. With the exception of WR, our models were very strong at reporting positives. The defensive position groups had lower sensitivity, indicating that the model was heavily over predicting positives. The offensive position groups had higher specificity. WRs had very similar accuracy, sensitivity, and specificity.

```
predictions_Dline  0  1
                0  4  9
                1 15 51

               Accuracy : 0.6962
                 95% CI : (0.5825, 0.7947)
    No Information Rate : 0.7595
    P-Value [Acc > NIR] : 0.9232

                  Kappa : 0.0678

 Mcnemar's Test P-Value : 0.3074

            Sensitivity : 0.8500
            Specificity : 0.2105
         Pos Pred Value : 0.7727
         Neg Pred Value : 0.3077
             Prevalence : 0.7595
         Detection Rate : 0.6456
   Detection Prevalence : 0.8354
      Balanced Accuracy : 0.5303

       'Positive' Class : 1
```

[1] D_line Random Forest Matrix

```
predictions_DBs  0  1
              0  9  8
              1 21 53

               Accuracy : 0.6813
                 95% CI : (0.5753, 0.7751)
    No Information Rate : 0.6703
    P-Value [Acc > NIR] : 0.46065

                  Kappa : 0.1897

 Mcnemar's Test P-Value : 0.02586

            Sensitivity : 0.8689
            Specificity : 0.3000
         Pos Pred Value : 0.7162
         Neg Pred Value : 0.5294
             Prevalence : 0.6703
         Detection Rate : 0.5824
   Detection Prevalence : 0.8132
      Balanced Accuracy : 0.5844

       'Positive' Class : 1
```

[2] DB Random Forest Matrix

```
predictions_RB  0  1
             0  9  4
             1 11 19

               Accuracy : 0.6512
                 95% CI : (0.4907, 0.7899)
    No Information Rate : 0.5349
    P-Value [Acc > NIR] : 0.08356

                  Kappa : 0.2825

 Mcnemar's Test P-Value : 0.12134

            Sensitivity : 0.8261
            Specificity : 0.4500
         Pos Pred Value : 0.6333
         Neg Pred Value : 0.6923
             Prevalence : 0.5349
         Detection Rate : 0.4419
   Detection Prevalence : 0.6977
      Balanced Accuracy : 0.6380

       'Positive' Class : 1
```

[3] RB Random Forest Matrix

```
predictions_LB  0  1
             0  4  2
             1 12 35

               Accuracy : 0.7358
                 95% CI : (0.5967, 0.8474)
    No Information Rate : 0.6981
    P-Value [Acc > NIR] : 0.33282

                  Kappa : 0.2382

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.9459
            Specificity : 0.2500
         Pos Pred Value : 0.7447
         Neg Pred Value : 0.6667
             Prevalence : 0.6981
         Detection Rate : 0.6604
   Detection Prevalence : 0.8868
      Balanced Accuracy : 0.5980

       'Positive' Class : 1
```

[4] LB Random Forest Matrix

```
predictions_WR  0  1
             0 22 11
             1 10 25

               Accuracy : 0.6912
                 95% CI : (0.5674, 0.7976)
    No Information Rate : 0.5294
    P-Value [Acc > NIR] : 0.004898

                  Kappa : 0.3813

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.6944
            Specificity : 0.6875
         Pos Pred Value : 0.7143
         Neg Pred Value : 0.6667
             Prevalence : 0.5294
         Detection Rate : 0.3676
   Detection Prevalence : 0.5147
      Balanced Accuracy : 0.6910

       'Positive' Class : 1
```

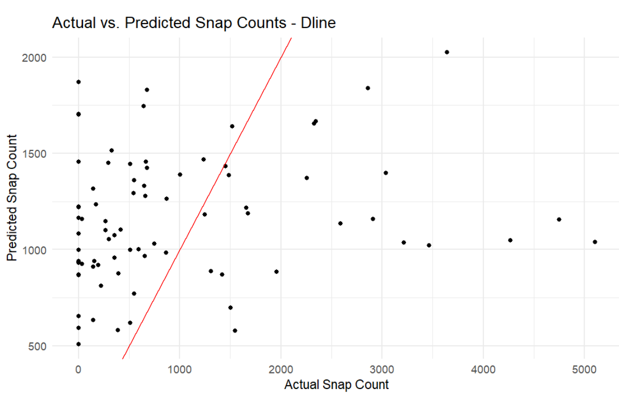
[5] WR Random Forest Matrix

Next we ran a linear regression model on our five position groups. Each regression model proved to be significant with low R squared numbers. For the defensive backs, the 40 yard dash had three star significance, while weight and shuffle time had one star. Linebackers showed two

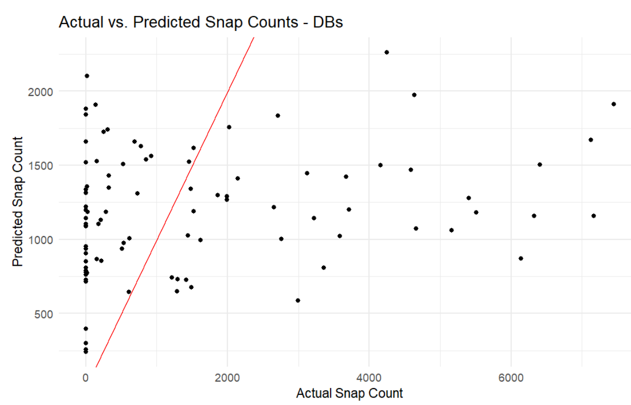
[1] 70% Accuracy, 85% Sensitivity, 21% Specificity
[2] 68% Accuracy, 67% Sensitivity, 30% Specificity
[3] 65% Accuracy, 83% Sensitivity, 45% Specificity
[4] 74% Accuracy, 95% Sensitivity, 25% Specificity
[5] 69% Accuracy, 69% Sensitivity, 69% Specificity

star significance with the 40 yard dash. Defensive line had a two star significance with height and one star significance with bench press. Running backs had a positive correlation with three star significance. They had negative one star correlations with height, 40 yard dash, and bench press. Finally, wide receivers had a two star significant relationship with the shuttle and one star with the 40 yard dash.

We then calculated the RMSE for each position group. The running backs had the lowest at 1016 and the defensive backs had the highest at 2027. We then plotted the predicted results vs the actual results. Overall the predictions were not very accurate.

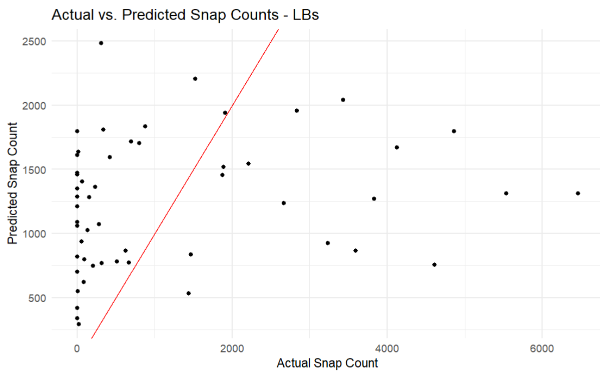


[6] Dline Linear Regression

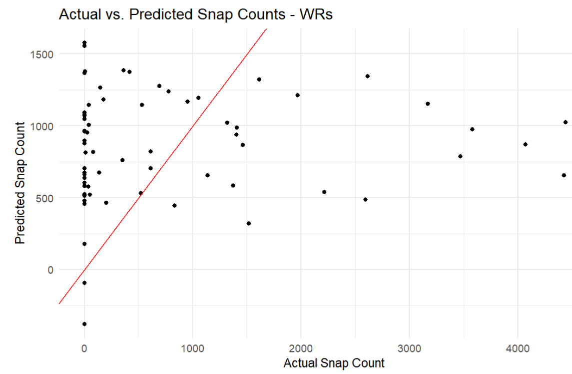


[7] DB Linear Regression

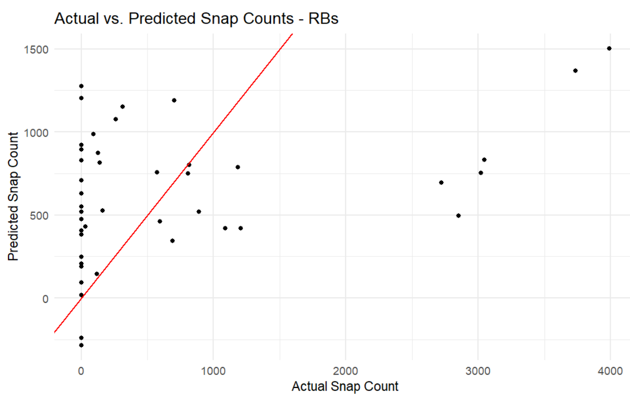
[6] RMSE: 1187.8304993586
[7] RMSE: 2027.76925039464



[8] LB Linear Regression



[9] WR Linear Regression



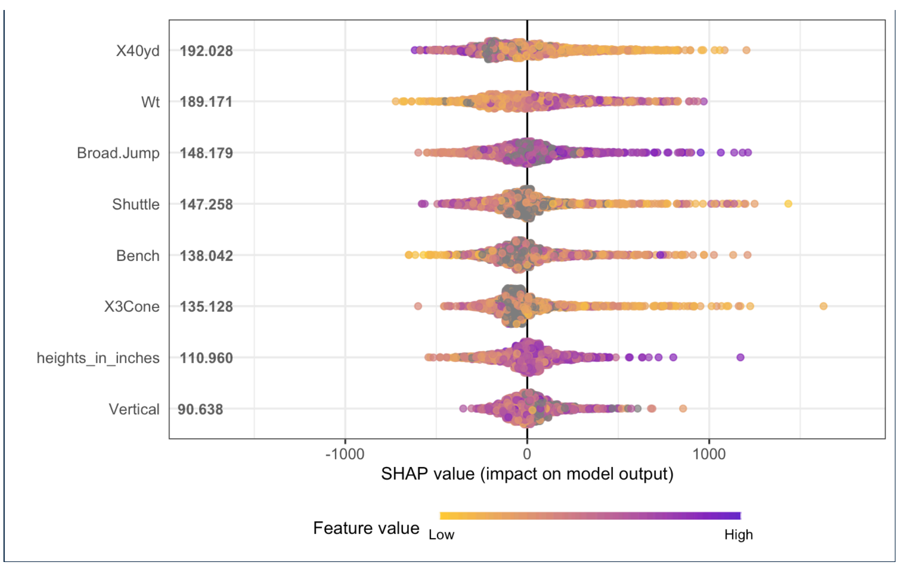
[10] RB Linear Regression

When using XGBoost, we elected to run the model for all positions combined, and then individually for each of our elected positions. Our model for each of the positions combined was the best overall predictor with an RMSE of 11.45, meaning that the predicted values differed from the actual values by an average of 11.45. The next best models were for Running Backs, Wide Receivers, and Defensive Linemen with respective RMSEs of 26.17, 28.41, and 35.07. Seeing as the actual snap counts for these positions reached a ranged from 0 to over 4,000, being off by 26-35 indicates that the model is fairly accurate. The models for Linebackers and Defensive Backs were much less accurate, with respective RMSEs of 192.59 and 286.51. While

[8] RMSE: 1636.34854569681
[9] RMSE: 1284.86604513386
[10] RMSE: 1016.85649825673

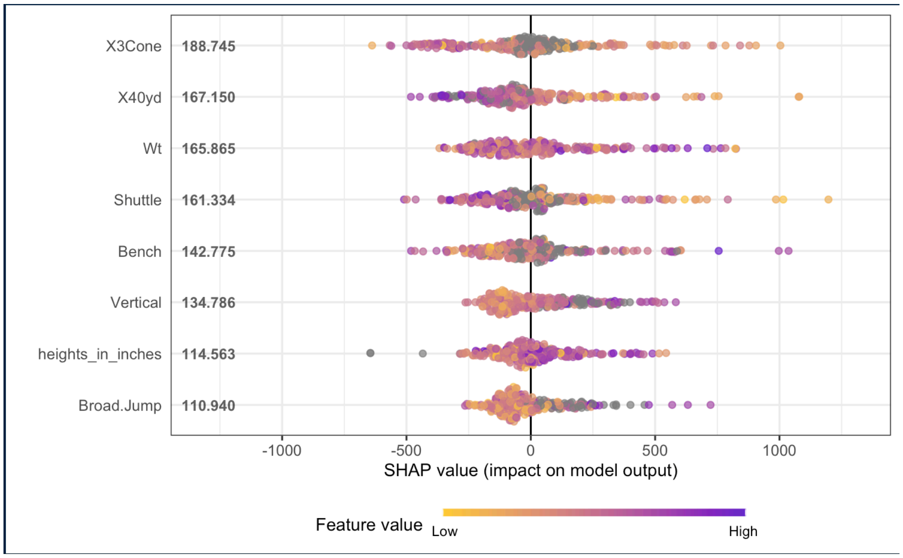
still decent considering snap counts reached over 4,000, these models were still much less accurate than their counterparts.

In addition to XGBoost, we created SHAP graphs to analyze the most influential combine events to the model. Focusing on our most accurate models, the top three predictors of the overall model were the 40yd dash, the weight of the player, and the player's broad jump. A faster 40yd dash was indicative of more career snap counts, as was a moderate weight and a farther broad jump. For Wide Receivers, the top three were the 3 cone drill, the 40yd dash, and the weight of the player. A faster 3 cone was indicative of more career snap counts, as were a faster 40yd dash and a moderate to heavier weight. For Defensive Linemen, the broad jump, the player weight, and the shuttle run were most influential. A farther broad jump, moderate weight, and faster to moderate shuttle were indicative of more career success. NAs did not seem to affect the model, as even Combine data with NAs showed to be influential in the SHAP graphs.

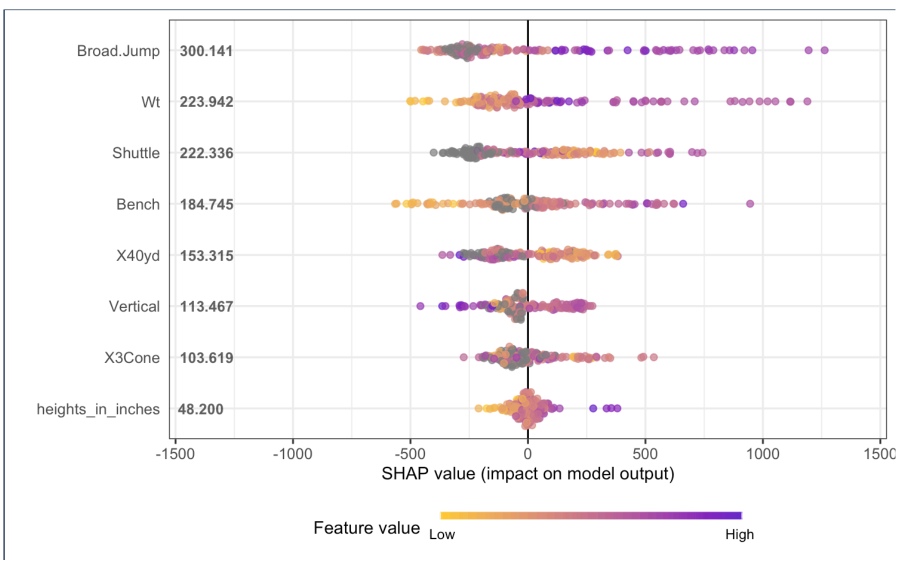


[11] SHAP graph for all positions

[11] Faster 40yd, moderate weight, and longer broad jump are most important in predicting more career snaps for the general player.



[12] SHAP graph for Wide Receivers



[13] SHAP graph for Defensive Linemen

Overall, from our three models we were able to assess the most important combine events in predicting player success for each position using XGBoost. XGBoost was also fairly accurate

[12] Faster 3 cone shuttle, faster 40yd, and moderate weight are most important in predicting more career snaps for Wide Receivers.
[13] Farther broad jump, moderate weight, and faster to moderate shuttle are most important in predicting more career snaps for Defensive Linemen.

in predicting career snap counts. Our random forest model was less successful in predicting career snap counts, and our linear regression even less so.

IV. Discussion and Conclusion

Our analysis provided useful insights into how NFL Combine performance relates to career success. The XGBoost model proved to be the most accurate, especially for positions such as Running Backs, Wide Receivers, and Defensive Linemen. Our overall XGBoost model produced the lowest RMSE at 11.45. In comparison, the Random Forest and linear regression models did not perform as well and the accuracy was far inferior. Furthermore, The SHAP analysis helped us identify which Combine events were the most predictive of career success. The 40-yard dash lived up to the hype, proving to be a strong indicator across all positions, with faster times leading to more career snaps. We also saw position-specific trends, such as the 3-cone drill being a strong predictor for Wide Receivers and the broad jump for Defensive Linemen. These findings demonstrate similarities and differences as to what athletic traits are important for each position. Speed shows to be a consistent factor, yet events that test change of direction or explosiveness show varying results across positions.

Although we dealt with missing data due to players opting out of certain events, this did not have a major impact on the accuracy of results. This demonstrates that even partial participation at the combine can still produce insightful results on a player's NFL potential.

In conclusion, our study shows that NFL Combine results, particularly in key events like the 40-yard dash and position-specific drills, can predict a player’s career success in the league. However, the accuracy of these predictions varies depending on the position. Moving forward, our findings can help NFL organizations make more informed decisions during scouting and drafting. Future research could build on this by combining Combine data with other factors such as college performance or psychological testing to get an even clearer picture of what leads to a successful NFL career.