# Predicting NFL Draft Pick Selections

*Based on Combine Metrics*

**Introduction**

Every year the National Football League (NFL) draft is one of the most anticipated events within the sporting industry. Last year, the 2024 draft increased its live audience to around 12.1 million viewers (Islam). Clubs have the opportunity to build on existing rosters or draft new talent to start creating the foundation of their team both offensively and defensively. Each year however, teams also run into the same dilemma in that there is a possibility that drafted players might be over/undervalued and therefore either are “over drafted” or “under-drafted” based on their respective round. Because of this, GMs often are subject to criticism when it comes to decision making, especially in earlier rounds (Robinson).

The NFL combine itself can be of the propagators of these issues. An individual can raise or lower their draft stock based solely on their performance within the NFL combine and how scouts perceive either their athletic ability or inability. As such, the NFL combine carries weight into where potential draft picks will be selected later. Lately the use of “RAS or relative athletic scores” (Denham), has been used to help summarize a player’s overall combine athletic ability compared to the rest of their respective fields. This gives teams more insight into what a potential player’s rank is out of their class athletically.

The NFL draft is heavily invested in by not only NFL teams, but also TV networks such as ESPN and the NFL Network. The 2023 draft brought in 35 hours of ESPN live programming and over 75 hours of NFL Network coverage (Greenberg, McCarthy). With the investment of capital in both the combine and the NFL Draft, this report aims to take a central focus on predicting draft position based on the athletic measurements and features present within the combine. The goal of the following predictive regression models will be to accurately assign a “Pick” number to individuals both attend the combine and are selected in the draft.

**Related Work**

Teams and organizations outside of the National Football League have investigated researching the topic of predicting a players’ draft slot. Mock drafts, which predict where players will be selected are released frequently, and 2025 mock drafts are already being created on CBS’ website (Trapesso). News companies such as *Sports Illustrated*, *Pro Football Focus*, and *USA Today* already combine traditional reporting with updated drafted standing based on combine metrics (Easterling, PFF,Middlehurst-schwartz). *HarvardSports* built multiple regression models in 2015 attempting to correctly predict NFL draft picks purely from combine data (Lotter). So, with that in mind, the following analytical process in this report shouldn’t be considered a novel concept.

**Data Description**

There were two data sets that were loaded into R environments for our model. The NFL Combine data itself was obtained through a *Kaggle* user’s repository and contained 4,741 player rows of 9 separate athletic measurements and features for every NFL combine performer from 2010 to 2023 alongside said player’s: name, position, school, draft round, draft pick, and year. The collegiate statistics data frame was part of an R package called CFB Fastr, which included 12,830 player rows of 54 offensive and defensive statistics for the selected seasons played (for this exploration, 2010-2023 to match the combine data) as well as each player’s: name, school, conference, year and ID. Each of these datasets was cleaned in different ways. The combine data contained duplicate player entries, so observations had to be grouped by unique player names. It also then had to be formatted correctly. Players’ heights were given in terms of feet and inches i.e. “6-3”, so a *stringr* method and conversion formula was used to convert this to total height in inches. Data types were all changed to numerical, and “Round” was excluded for the 2nd time running the *XGBoost* model’s prediction of specific pick number since “a draft round” would have the tendency to cluster picks into specific groups naturally on its own, and usually isn’t included in combine measurements as player rounds are still not given facts at the time. That said, it was included in the first version of the model because hypothetical “mock drafts” do exist so while that isn’t a part of combine data, it can still lead to interesting results using combine measures as almost additional data to supplement player predictions. Year was also later excluded do to it potentially causing noise mixed in with other numerical features.

Likewise, the collegiate statistics data frame had to be formatted correctly. Statistics had to be scaled across the data frame because some applied to unique position groups leaving large numerical discrepancies across the board. Then “Player ID” had to be dropped since it would introduce noise into future models alongside player name since it was a character data type. Conferences were kept as factors since they and years both represented factors in the random forest scenario and “team” needed to either bet dropped completely or converted to a binary factor representing dummy columns for each individual collegiate team. Lastly statistical features were then changed to numerical data types.

On first attempting to join both tables it was quickly recognized that after filtering out players who had not been drafted from the combine data, and duplicates, there were only 329 rows for the combined 73 features (some of which also contained duplicated). It’s not known currently what caused this sharp decline from the initially expected 3,008 players in the clean version of the combine data set, but it is important to note ahead of time that any modeling based on such a small sample size would have had vastly larger chances of overfitting to the training and validation data, and thus become worse when predicting for future NFL players. With that in mind, it was determined that the projected pick would be predicted on the combine dataset only since it did represent the largest availability of observations to train models.

**Methods**

The methods primarily used in this exploration were an XGboost model and a Random Forest model. The XGboost model was built off the combine data alone because of its ability to take purely numerical data and create a regression predictive model. In this case, predicting a “Draft Pick number” based on all the combine numerical features. It was decided to use a Random Forest model for the purpose of comparison to the XGBoost model. Having said that, because of the lack of tuning within the random forest model, it was not a one-to-one comparison. In other words, the XGBoost model was naturally going to outperform the random forest model because of the extensive tuning.

**Results**

The results centered around the combine data. The college football statistics were useful on their own; however, the dataset lacked the response variable, ‘Pick’ - hence why the college football dataset was omitted from analysis.

The XGBoost model was utilized because of its effectiveness in optimizing accuracy. The initial XGBoost model calculated an RMSE of 11.715(for the model including round) and ~59(for the model not including Round). Furthermore, when plotting the importance of the variable utilizing the xgb.importance() function we found that the predictor, ‘round’ was the most important variable in predicting the response, ‘Pick’ or pick selection. That being said, the variable ‘round’ created uncertainty within the group. To explain, the group was unsure if the variable, ‘round’ was the projected round a player would be selected, or if that variable represented the actual result/actual round the player was selected. If the variable represented the actual result, the data would naturally be biased because the data would be combining predicted results with actual results. Because of the uncertainty, ‘round’ was removed for one of the XGBoost models (the code reflects that it was negatively selected in the “clean\_combine\_boost\_data” just prior to data partitioning). To see the impacts of the model with “round”, one need only leave it out of the negative select statement and change the “lapply” line from 1:11 to 1:12, alongside making the train and test matrices 1:11 instead of 1:10 or vice versa for the model without “round” included (to account for the additional column Lines).

The variable importance plotting displayed insight that the top 4 features driving the XGboost’s prediction after “Round” were: 3 cone drill, weight, 40-yard dash, and shuttle. From this analysis, it can be seen that the model is focusing on athletic events that center around how agile a player was, how in shape they were, and how fast they were. After plotting the variable importance, tuning was conducted. The first parameter tuned was the number of trees. Alike the other hyper parameters, finding the ideal number of trees to utilize is critical in establishing a peak model. The code written found that 51 was the best iteration for the model with “Round” and 31 for the model without.

Following the number of trees was tuning the depth and weight. We tested multiple depths and weights through a for-loop and found that a depth of 3 and weight of 1 was ideal for the round model and a depth of 5 with a child weight of 15 for the non-round model (Figures 3 and 9).

A chart of different colors

Description automatically generated

Another hyper parameter addressed was Gamma. Gamma works to obtain very little loss reduction; smaller gamma values would reflect an aggressive model (XGBoost). The code written found values close to .20 were ideal for the round model but the non-round preferred a gamma of .15.

Subsample and column sample were also tuned. Subsample is important in building either a complex or simpler model. The importance of subsamples is that the data will be overfit or underfit depending on the subsample (Medium). When visualizing the subsample and column sample, the model using round aligned with both values being 1. That however was not the case for the model not using round as it had its lowest RMSE value as a combination of .9 subsample and .7 column sample (Figures 4 and 10).

A colorful chart with white text

Description automatically generated with medium confidence

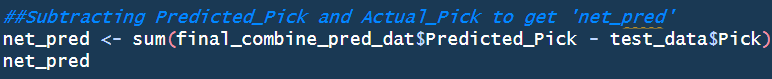
The learning rate or ETA was also inspected. After plotting multiple learning rates, it was concluded that 0.3 was actually the preferred eta for both models since it achieved the same RMSE level as 0.1 but learned faster (Figure 5 and 11).

A graph of different colored lines

Description automatically generated

Finally, once tuning was completed, the last XGBoost model was built.

Because the response variable was non-binary, confusion matrix, specificity, accuracy and additional metrics were not computed. The metrics that were utilized to evaluate the performance of these models consisted of RMSE, residual sum of squares alongside an improvised metric called, ‘net\_pred’. ‘Net\_pred’ was the sum of all the residuals, or the sum of the difference between the predicted pick and the actual pick.



**Discussion**

Our final XGBoost Model including “Draft Round” was able to produce a RMSE of around 9.29 unlike the model not including “Draft Round”, which finished with an RMSE value of 59.99. While RMSE is valuable, a metric that both evaluated the performance of the model and was easy to interpret was ‘net\_pred’. The net\_pred the final XGBoost model with “Round” produced was within the high abs(200-300’s) range. This means that our model was about 200-300 selections off in total. For comparison, the random forest model was approximately in the abs(700)’s range(before knitting). The significance of a lower ‘net\_pred’ is that the model is a more accurate total predictor of the actual pick.

The interesting component about analyzing a model including “Round” alongside ones that purely use combine data or aren’t as strong models is that “Round” tends to naturally cluster players towards the right direction, while some of the combine metrics alone leave more to be desired in terms of purely predicting position based on athletic measures. What this primarily means is that predicting a player’s draft position purely on their NFL combine data is incredibly challenging alone. That said, if combine data is paired with say a mock draft projection, already grouping players by talent into respective draft rounds, one would be able to train on the “Round” XGboost model and then predict using the expansive “mock draft” projections analysts churn out on a weekly basis. The addition of player statistics alongside these combine data models, would only increase the predictive power of future draft projections. With more tuning alongside statistics, scouts and GMs can purely focus on some of the more intangible aspects of draft evaluation, such as reviewing film, while machine learning models such as these help to ease the overall process.

**Conclusion and Future Work**

The problem this team framed, centered around accurately predicting pick selection. It was found that utilizing an XGBoost model was ideal for answering the question relative to a random forest model because of the lower RMSE and ‘net\_pred’. Through variable analysis, we found that ‘Weight’, the 3-cone drill,40-yard dash, and shuttle were important variables in predicting draft selection. Overall, the main XGBoost model was able to somewhat accurately predict pick selection as the sum of the residuals was in the 200 range. In other words, the Round XGBoost model was only around 200 selections off in total the actual pick. Having said that, for future work, it is possible to consider other metrics such as college statistics and ‘film’ to obtain the most accurate player valuation and pick selection.

Contribution

Josh obtained both datasets, cleaned them, wrote the XGBoost model code as well as the Introduction, Data Description, and Methods used section of the report. Ryan performed the random forest model and also wrote the results, discussion, and future work components of the report. Everyone contributed to the presentation and analysis.

Bibliography

* Shaw, Thomas S. *NFL Combine Performance Dataset*. Kaggle, <https://www.kaggle.com/datasets/thomassshaw/nfl-combine-performance-dataset>. Accessed 23 Sept. 2024.
* "CFB FBS Reference." *CFB Fastr*, Sports Data Verse, [https://cfbfastr.sportsdataverse.org/reference/cfbd\_stats.html. Accessed 23 Sept. 2024](https://cfbfastr.sportsdataverse.org/reference/cfbd_stats.html.%20Accessed%2023%20Sept.%202024).
*  Relative Athletic Scores. "The Importance of RAS in Football Scouting." RAS.football, <https://ras.football/>. Accessed 24 Sept. 2024.
*  Denham, Joey. "NFL Combine: Offensive Winners and Impressive RAS Scores." Stadium Rant, 6 Mar. 2024, <https://www.stadiumrant.com/post/nfl-combine-offensive-winners-and-impressive-ras-scores>. Accessed 24 Sept. 2024.
*  Islam, Arif. "2024 NFL Draft: Round One Viewership and Attendance in Detroit Surpass Expectations." SportsPro, 29 Apr. 2024, <https://www.sportspromedia.com/news/2024-nfl-draft-round-one-viewership-attendance-detroit-espn-abc/>. Accessed 24 Sept. 2024.
*  Robinson, Jacob. "Six Draft Day Mistakes NFL GMs Make, Plus Ideal Fantasy Football Landing Spots." The New York Times Athletic, 25 Apr. 2024, <https://www.nytimes.com/athletic/5445068/2024/04/25/six-draft-day-mistakes-nfl-gms-make-plus-ideal-fantasy-football-landing-spots/>. Accessed 24 Sept. 2024.
*  Greenberg, Doug, and Michael McCarthy. "Nobody Wanted to Show the NFL Draft. Now It Owns TV." *Front Office Sports*, 1 May 2024, <https://frontofficesports.com/nobody-wanted-to-show-the-nfl-draft-now-it-owns-tv/>. Accessed 24 Sept. 2024.
*  Trapesso, Chris. "2025 NFL Mock Draft: Panthers Move On from Bryce Young; Shedeur Sanders, Quinn Ewers Go to Surprising Teams." CBS Sports, 19 Aug. 2024, <https://www.cbssports.com/nfl/draft/news/2025-nfl-mock-draft-panthers-move-on-from-bryce-young-shedeur-sanders-quinn-ewers-go-to-surprising-teams/>. Accessed 24 Sept. 2024.
* Easterling, Luke. "2024 NFL Draft Rankings: Running Back – Trey Benson, Jonathon Brooks, Jaylen Wright." *Sports Illustrated*, 4 Oct. 2023, [www.si.com/nfl/draft/news/2024-nfl-draft-rankings-running-back-trey-benson-jonathon-brooks-jaylen-wright. Accessed 6 Oct. 2024](http://www.si.com/nfl/draft/news/2024-nfl-draft-rankings-running-back-trey-benson-jonathon-brooks-jaylen-wright.%20Accessed%206%20Oct.%202024).
* PFF. "2024 NFL Combine Results." *Pro Football Focus*, 2024, [www.pff.com/news/draft-2024-nfl-combine-results](http://www.pff.com/news/draft-2024-nfl-combine-results). Accessed 6 Oct. 2024.
* Middlehurst-Schwartz, Michael. "NFL Combine Winners and Losers from 2024 Draft: Xavier Worthy Impresses." *USA Today*, 4 Mar. 2024, [www.usatoday.com/story/sports/nfl/draft/2024/03/04/nfl-combine-winners-losers-draft-2024-xavier-worthy/72835664007/](http://www.usatoday.com/story/sports/nfl/draft/2024/03/04/nfl-combine-winners-losers-draft-2024-xavier-worthy/72835664007/). Accessed 6 Oct. 2024.
* Lotter, Bill. "The Combine Actually Matters, Part 3: Predicting the Draft." *Harvard Sports Analysis Collective*, 19 Feb. 2015, harvardsportsanalysis.org/2015/02/the-combine-actually-matters-part-3-predicting-the-draft/. Accessed 6 Oct. 2024.

A graph showing a line of blue dots

Description automatically generated

**Figure 1:** First XGBoost Model with round included as a feature.

A graph with numbers and letters

Description automatically generated with medium confidence

**Figure 2**: Displays the variable importance graphics of the XGBoost model including round.

A chart of different colors

Description automatically generated

**Figure 3**: Displays the heatmap of Tuning depth and child weight for the XGboost with round.

A colorful chart with white text

Description automatically generated with medium confidence

**Figure 4**: Displays the heatmap of tuning column sample and subsample for the XGBoost with round.

A graph of different colored lines

Description automatically generated

**Figure 5**: Displays the ETA curve for tuning the XGboost with Round model.

A screenshot of a computer

Description automatically generated

**Table 1**: Displays the first 10 predicted vs actual pick selections for the XGboost with round.

A graph with blue dots

Description automatically generated

**Figure 6**: Visualizes the Final Predicted vs Actual Pick selections by round for the XGBoost model with round.

A graph showing a diagram of values

Description automatically generated with medium confidence

**Figure 7**: Displays the first XGboost model when “round” wasn’t used as a feature.

A graph of a number of objects

Description automatically generated with medium confidence

**Figure 8**: Visualizes the important variables within the XGboost model not using round.

A chart with different colors

Description automatically generated

**Figure 9**: Displays the heatmap of Depth and Child Weight tuning for the XGboost model without round.

A colorful squares with numbers

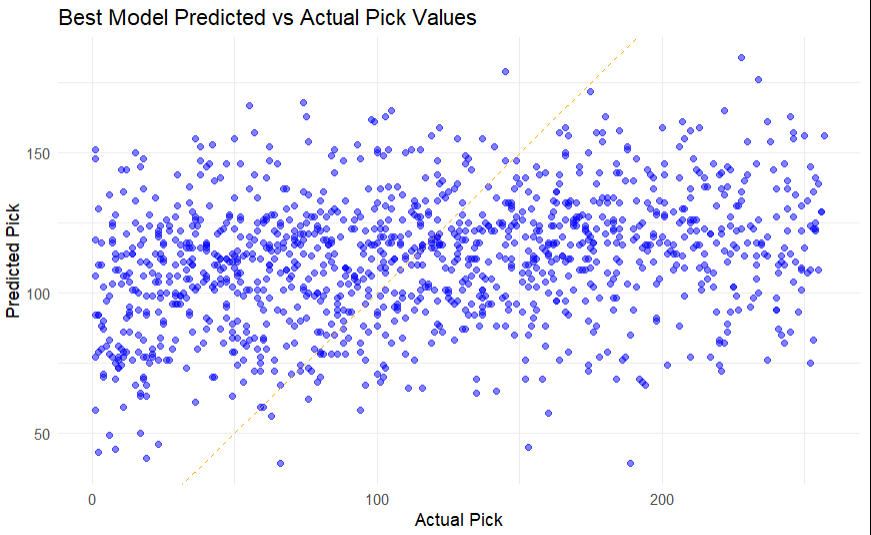
Description automatically generated with medium confidence

**Figure 10**: portrays the heatmap of Column Sample and Subsample tuning for the XGboost model without round.

A graph of different colored lines

Description automatically generated

**Figure 11**: Shows the ETA Tuning curve for the XGBoost model without round.



**Figure 12**: Displays final Predicted vs Actual Picks for the XGBoost model without round.

A screenshot of a black screen

Description automatically generated

**Table 2**: Shows the first 10 predicted vs actual picks for the XGboost model without round.