

[NFL Mock Draft 2022: Tyreek Hill, Deshaun Watson, Davante Adams trades shake up Round 1 for Chiefs, Packers and Texans | Sporting News](https://www.sportingnews.com/us/nfl/news/nfl-mock-draft-2022-davante-adams-deshaun-watson-nfl-free-agency/rmoj8obwo8zv5k92srdlncb8)

The NFL Draft is a momentous event for American Football fans and players. Since 1936, the NFL Draft has been the event in which fresh collegiate players are recruited into the league. Currently, the draft consists of 7 rounds with each round having around 32 picks. For a player, the goal is to be selected by a team within these 7 rounds, but how do these teams determine what players should be drafted? In order to determine this, teams look at raw talent measurements taken during the NFL Combine, technical ability displayed in drills and college gametape, and many other indicators of a potentially elite player.

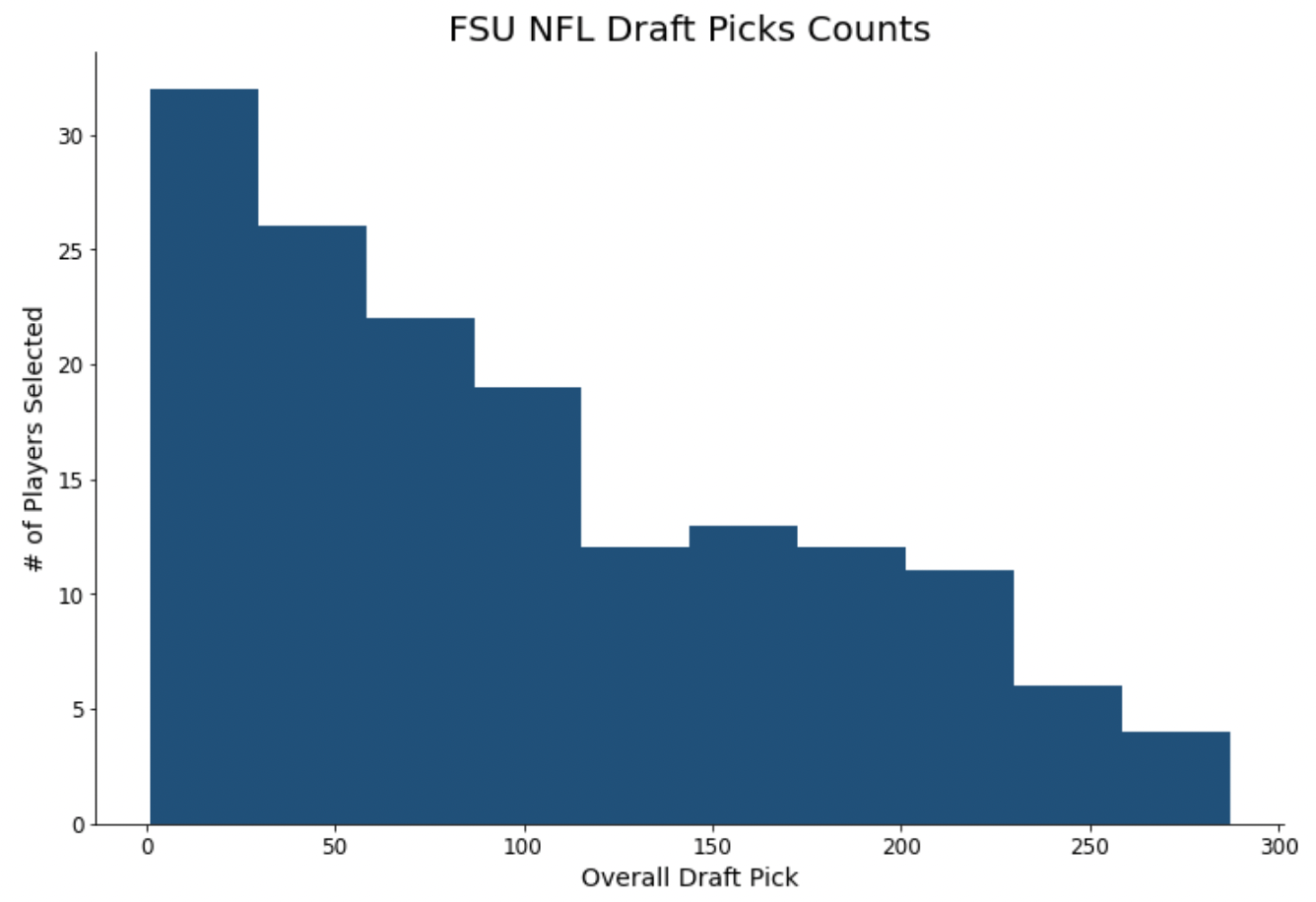
As mentioned above, the NFL Combine gives players an opportunity to showcase their raw athletic abilities through several events and measurements. These events include the 40-Yard Dash, 20-Yard Shuttle Run, Vertical Jump, Broad Jump, and Bench Press. Prospect’s height, weight, hand size, and arm length are also taken during the event.

Our group was fascinated by this process and were interested in the idea of creating a model that is capable of predicting the round and/or overall pick for each player in a given year’s NFL Draft based on the player’s combine performance. To accomplish this, we’re working with a dataset from [Kaggle](https://www.kaggle.com/datasets/toddsteussie/nfl-play-statistics-dataset-2004-to-present?select=players.csv) that includes information on over 10,000 current and previous NFL players. For our purposes, we’re focusing on the data that concerns the NFL Draft and NFL Combine in particular. This data contains information such as a player’s name, college, weight, height, round they were selected, pick they were selected, and all of their results from the NFL Combine.

To start cleaning our data, we dropped unnecessary columns in both datasets. Then, we merged both datasets on the ‘playerID’ column in order to only keep players that we had combine and draft data for. We also filled all null values with an empty space (“ “) or a zero depending on the data type of the column.

In order to get a better feel of our data, we started our project by exploring some of the interesting situations that arose in certain columns.

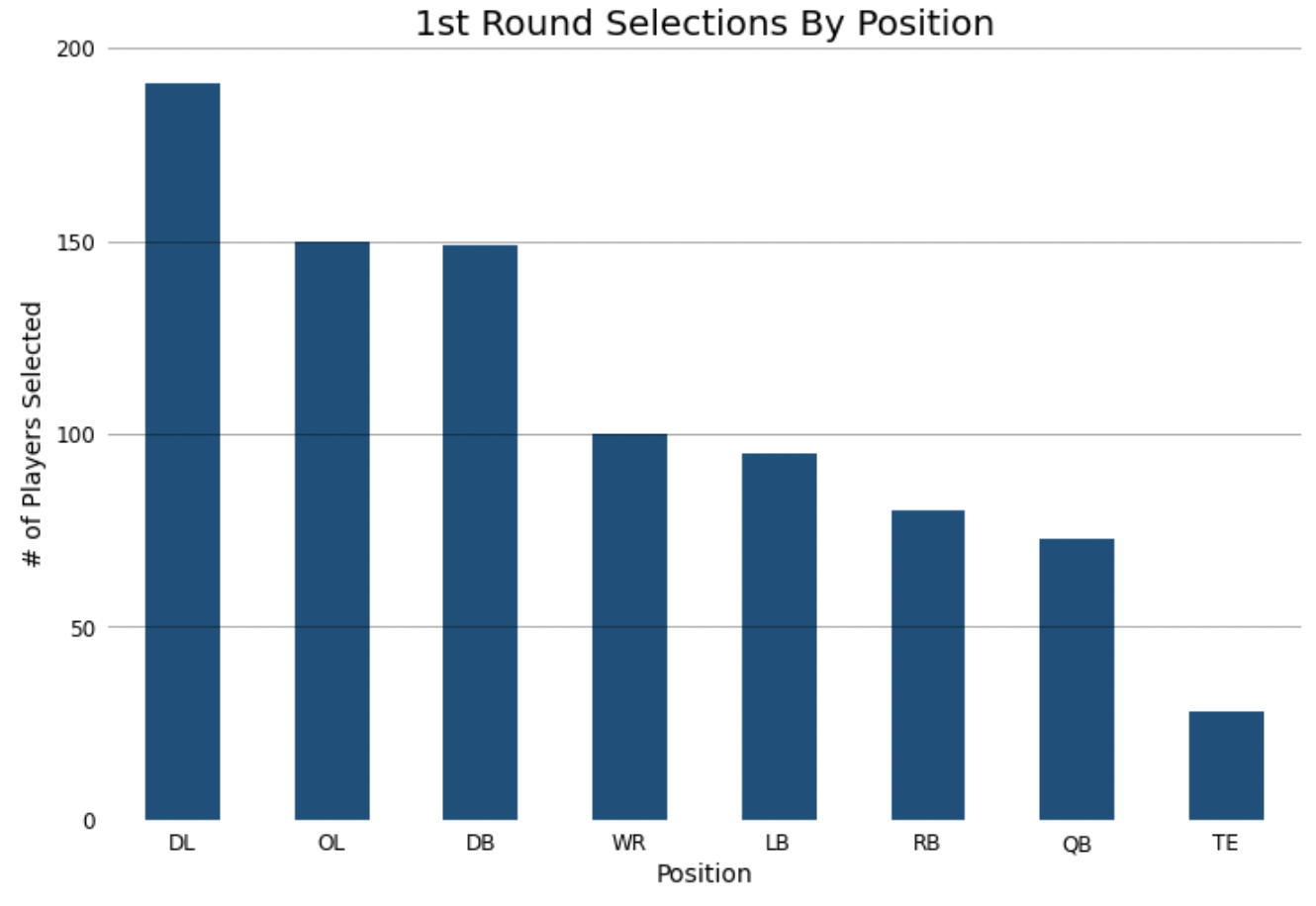
First, we looked at the college that each prospect attended to see if there was any bias in how many players from each school are drafted. It turns out that some schools are drafted way more than others; this makes intuitive sense because only certain schools can compete at an elite level, and these schools’ players are most likely the ones that will be best prepared to play professionally and thus be drafted. We found that the most drafted college in our data was Florida State University. *Figure 1* shows us that over 150 prospects were drafted out of Florida State, with a large majority of those selections being in the early rounds of the draft.



*Figure 1*

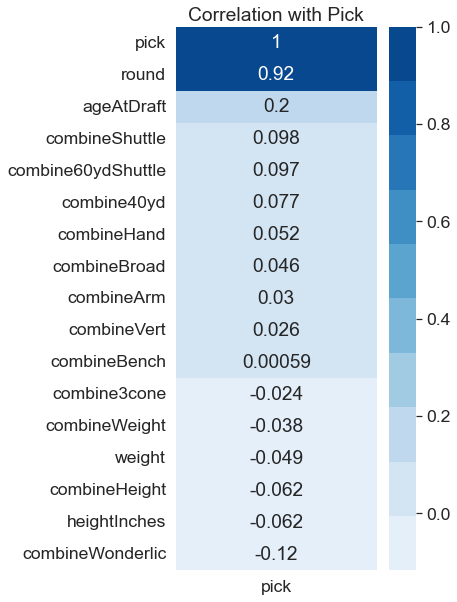
\

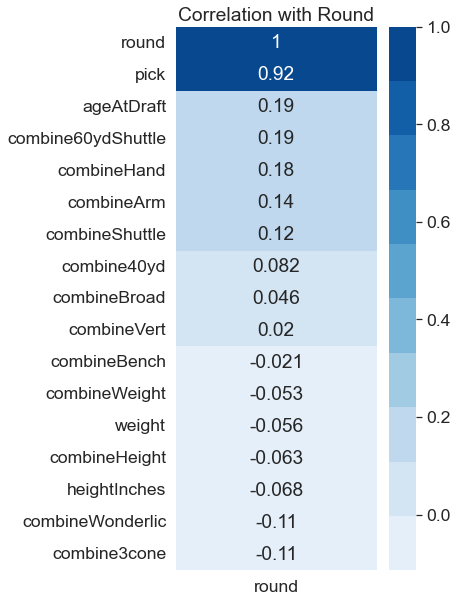
Next, we decided to look at which positions were most commonly drafted in the first round. On the surface, one would assume that the players that get paid the most would be drafted the most early-on. However, *Figure 2* shows us that players from the defensive and offensive lines are the first and second most commonly selected in the first round, respectively. Interestingly, very few quarterbacks are drafted in the first round, but when they are, they are normally the first overall pick.



*Figure 2*

After manually looking at our data to determine what will be important for our models, we decided to see what Pandas’ built-in functions could offer. *Figure 3* and *Figure 4* show the correlation coefficients between the different variables in our data ane either Overall Pick or Round.

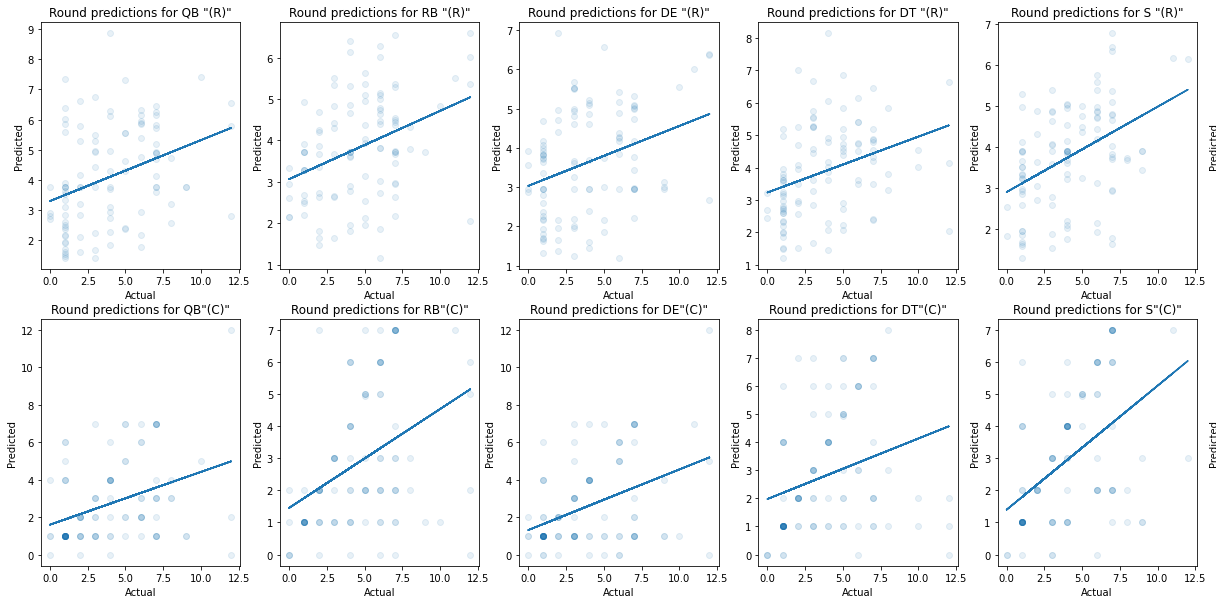




*Figure 4*

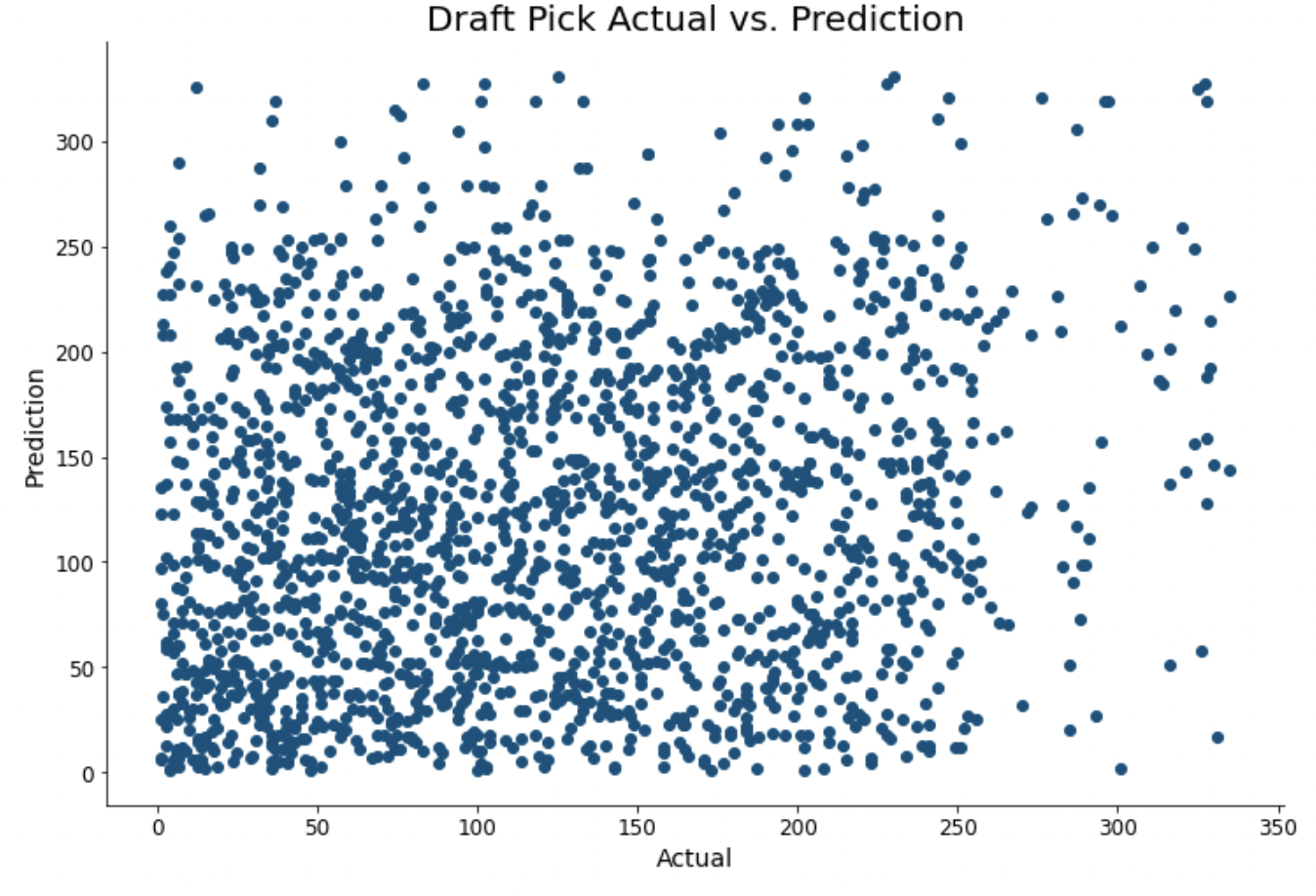
As we can see in both figures, the columns have very little correlation to overall pick and round. This could represent a problem for us when creating our model.

Once we had a decent feel for our data, we began building our models. Our group decided to use Nearest Neighbors, Random Forest, Decision Tree and XGBoost regressors and classifiers. The columns that we used in our models were: 'combine40yd', 'combineVert', 'combineBench', 'combineShuttle', 'combineBroad', and 'combine3cone'. Each model was run for each of the position groups and GridSearchCV was used to ensure that the best hyperparameters were used. None of the models showed strong linearity between actual values and the predicted values, hence we decided to average the predictions for all the regression models and take the mode of the classification models in an attempt to optimize our model. The scatter plots slightly improved, which can be seen in *Figure 5*.



*Figure 5*

Our second model is a Decision Tree Regressor that uses all measurements and events that take place in the modern NFL Combine. The model takes these as inputs and then splits the data into training and testing sets to predict the overall pick for each prospect. The results of running this model can be seen in *Figure 6.* Our model is way off on some predictions and very accurate with others; this prompted us to wonder how the model was operating on a case-by-case basis. Our model fails to take into account a prospects technical ability and other factors that can play a key role in how high they are selected. For example, the predictions seen in the upper left quadrant of our plot are players with high technical abilities but poor combine results, whereas the bottom right quadrant is the opposite. When our model is very accurate with its prediction, it tends to be the case that those are the players with the ideal combination of technical skills and raw athletic ability.



*Figure 6*

The most glaring problem with our models is the lack of technical ability. We’ve seen that raw athletic ability is not the only important factor when it comes to how high a player is drafted. A great example of this comes when viewing the record holders for several combine events; of the 5 events, only 2 of the record holders were taken in the second round or higher. This idea, when combined with several nuances of the draft and combine, creates a very complex situation to create a model for. However, we do believe that if a model is given access to measures of technical ability, then it will perform much better than what we have presented above.