Libby Czarniak and Emily Greco

Final Project- MA346

Professor Carter

December 3, 2020

**Introduction**

For our final project we built a dashboard to analyze NFL combine data and made regression models to predict NFL statistics from NCAA statistics. When deciding which players to draft, NFL coaches, managers, and scouts use a variety of statistics and assessments to make their final decisions. At the end of the day, though, that decision comes down to a player’s skill, character, and how well he fits with the team. Obviously, character and whether someone will mesh with a team is hard to quantitatively measure, so we found different measurements of skill that may help make draft decisions. These measurements include statistics from the NFL Combine and NCAA game statistics.

The NFL Combine, also known as the NFL Scouting Combine, is a week-long set of mental and physical tests that college players perform. The physical tests measure players’ strength, speed, and agility, for example, and allow NFL coaches, managers, and scouts to evaluate prospects in a standard way. These results can help make draft decisions. Our goal is to use past NFL Combine Statistics to provide an idea of how players performed at the Combine in the past and to make recommendations for what statistics NFL teams may want to look for from prospective players.

NFL Combine statistics are limited, however, in what they suggest about how players might perform in games. Therefore, we also found NCAA game statistics to try and predict how drafted college football players might perform in the NFL. More specifically, we set out to build regression models that predict certain NFL statistics for four different categories of positions. Ultimately, we found that using college game statistics are not very predictive of NFL statistics.

**Data Collection and Organization**

We found the dataset we used for our NFL Combine analysis on Kaggle (<https://www.kaggle.com/savvastj/nfl-combine-data>). This dataset has NFL Combine test statistics for players spanning the years 2000 to 2017. The dataset also has information regarding each of the players, including his name, position, height, weight, what year and round he was drafted in, which pick he was in that round, and the NFL team that drafted him. No cleaning was necessary prior to importing the data file. After importing the dataset as a Pandas dataframe, we realized that both players who had been drafted and players who had not been drafted were included. We decided to keep only the players who had been drafted to ensure we made the best possible recommendation for NFL coaches, managers, and scouts. At that point, we cleaned the dataset to convert all columns to their appropriate data types.

Given that the dataset contains many positions, we focused on only five of the more popular ones: quarterback, running back, defensive tackle, outside linebacker, and wide receiver. The dataset has statistics for six NFL Combine tests as well, so we analyzed the three more important ones for each position. After selecting the positions and their corresponding test statistics, we made a new dataframe for each position. These five data frames contained data for only the players at the position, height and weight, the Combine tests relevant to that position, and the draft round, pick, and year for each player. We used these throughout the rest of the project.

Building a dataframe for creating regression models required more preparation. We could not find a single dataset with both NCAA and NFL statistics for players, so we pulled datasets from two different sources on Kaggle. The [NFL Play Statistics dataset (primary)](https://www.kaggle.com/toddsteussie/nfl-play-statistics-dataset-2004-to-present) includes information on players and their game statistics spanning the 2004-2019 seasons. The [College Football Statistics](https://www.kaggle.com/mhixon/college-football-statistics) datasets include player and game information from the 2005-2013 NCAA seasons, but we used data from only 2008-2013. In order to create a dataframe sufficient for building models, two major steps were required: identifying which players existed in both datasets and matching a player’s college statistics to his NFL statistics.

In order to figure out which players existed in both the NFL and NCAA player datasets, we created dataframes for both and merged them. Preliminary exploration of the data indicated that several columns would be needed to join the dataframes, so we had to ensure that both datasets contained those columns. The NFL player dataframe had all the necessary columns, and each row represented an individual player. It contained all players from every season, so no preparation or cleaning was necessary. On the other hand, the NCAA datasets contained separate files with player information based on season. These datasets contained the same columns and had individual players represented by rows, so we concatenated, or joined, them together to create one dataframe and added team name to it. At that point, we created one large dataframe that contained players who were in both the NFL and NCAA dataframes, and we cleaned it to keep only the columns we wanted and ensure there was just one row for each player.

The NCAA dataset had game statistics separated into files based on season, but they all contained the same columns so we could easily join them together. There were more than fifty columns with statistics for different types of plays, so we kept only the columns that made sense to use for predicting statistics. Furthermore, we calculated cumulative statistics for players across every game they played in that existed in the dataframe and then put them on a per game and per play basis, which made the most sense for creating regression models. The NFL datasets contained different types of statistics in separate files, but they included every game from the 2004-2019 seasons. We turned each of these datasets into a dataframe with only the columns we needed for making models and merged them into one.

The NCAA statistics dataframe and the dataframe with NFL statistics had columns that allowed us to merge them with the dataframe containing player information. At that point, we had one large dataframe with rows representing individual players and columns representing player information, NCAA statistics, and NFL statistics. This dataframe allowed us to pull statistics for creating regression.

**Data Visualization on the Dashboard and Modelling**

Dashboard Link: <https://secret-gorge-55796.herokuapp.com/>

With our dashboard, we wanted to give the user an idea of how different positions performed in past NFL Combines. Users can provide input for the draft round, position, and Combine test they are interested in. This provides a visualization- side-by-side boxplots- of data from past Combines based on user input. They can also get summary statistics from past Combines and a recommendation for what statistics to look for from prospective players based on their input. Furthermore, the user can provide a Combine test statistic for a player they are considering drafting to get an idea of how he compares to past players.

For the modelling portion of the project, we used Sklearn’s machine learning, a Python package, for multiple linear regression. Since our goal is to help NFL coaches, owners, and scouts make informed decisions when they draft college players, we aimed to determine whether we can predict NFL statistics that measure skill for certain positions. To accomplish this goal, we built regression models that use college statistics aligned with that NFL skill to predict those NFL statistics. Prior to making the models, we used plots to determine whether there may be a linear relationship between the outcome and predictor variables. To build these models, we used Sklearn features that allowed us to split our data with the variables for the model into two partitions. One partition was used to build the model, and the other was used to test how useful the model was. Sklearn also has a feature that provides an R2 value for the model. R2 indicates how well the model explains the data, or how useful it is for predicting the response variable.

**Interesting Findings**

When performing some small analyses of the boxplots for the Combine statistics, we noticed a couple of interesting trends. We expect that statistics would be better for players drafted in the first few rounds and get worse for players drafted in later rounds. That was the case for most of the Combine tests for each position. For example, when examining the boxplots for the forty-yard-dash for running backs, there is a clear upward trend in the median time (see Exhibit 1). Running backs drafted in the first round have a much lower median time as compared to players drafted in other rounds. With the exception of the second round, the median time increases significantly for players drafted later on.

However, in other cases, there was not a big discrepancy between the statistics in each round. For instance, if we analyze the box plots for Wide Receivers’ shuttle statistics, we notice that the median time is pretty constant from the first to the seventh round. Given that players drafted early on are considered to be superior in skill, we might expect that their shuttle time would be significantly lower than times for players drafted in later rounds. This does not seem to be the case, however, and it supports the fact that these Combine tests do not measure game performance and other tangible factors NFL coaches, owners, and scouts look for in players.

Our most interesting findings come from our attempt to determine whether we can build a regression model that predicts NFL statistics using college game statistics. We focused on statistics for quarterbacks, running backs, wide receivers and tight ends, and defensive positions: defensive tackles, linebackers, outside linebackers, defensive ends, and defensive lineman.

For quarterbacks, we created a regression model to predict passing yards per NFL game using NCAA pass completion percentage, yards per completion, and passing touchdowns per game. These college statistics made the most sense to use based on our available data, and preliminary plotting indicated that they might not have a strong linear relationship with our variable of interest. The best model we created had all of these NCAA statistics as explanatory variables. However, given the model’s R2 value of .18, we concluded that we could not make a very useful prediction for NFL passing yards per game using these variables. While this is not a strong enough model to make a prediction, we can point out that pass completion percentage and yards thrown per completion have a bigger effect on the predicted passing yards per NFL game. It may be worth considering these statistics when assessing how well a quarterback may perform in the NFL.

When considering running backs, we strove to determine whether we can predict rushing yards per NFL game using a player’s NCAA yards per carry, rushing touchdowns per game, and fumbles per game. Our best model contained all these statistics, but it did not prove to be useful. When we fit the model to new data, its R2 value was -.59, which indicates that the model does not fit the data or predict the outcome of interest well. When we fit the model to the data we used to actually build the model, we got a stronger R2 value of .34. However, given that the model does not explain trends in new data, we cannot conclude that we are able to make an interesting prediction for this NFL statistic using the available college statistics. Commenting any further on the model is not warranted.

Wide receivers and tight ends typically have similar statistics, so we grouped them together to predict their reception yards per NFL game. For this specific statistic and these positions, the number of both reception touchdowns per game and yards gained per reception in college made the most sense to use as predictors (see Exhibit 2). The most useful model contained both of these college statistics. This model had an R2 value of .19 when fit to new data and an R2 value of .23 when fit to the data used to make the model. Again, the R2 values are not overly strong, so it seems unlikely that we can predict NFL reception yards per game using the college statistics at hand. Despite this, we will note that NCAA reception touchdowns per game have a bigger impact on the outcome of interest than yards gained per reception. An NFL coach or scout may want to look into this statistic when assessing players at these positions.

Lastly, we grouped together several defensive positions that are likely to have similar statistics and created a model to predict tackles per NFL game. Our dataframe contained several college statistics that could be useful for predicting this NFL statistic. But, the most useful model contained solo tackles, tackles for a loss, and sacks per college game along with yards lost on a sack by the player. Again, this model did not prove to be very useful for predicting tackles per NFL game. When fit to the data used to build it, the model produced an R2 value of .17 and an R2 of .14 when fit to new data. Although the model is not strong enough to make a prediction, it is worth noting that solo tackles and tackles for a loss in each NCAA game had the biggest impact on the outcome of interest. We will note that these could be helpful statistics when measuring a player’s skill and making a draft pick.

**Conclusion**

We conclude that, overall, we cannot predict players’ performance in the NFL well with college statistics. The college statistics we found were the best measure of skill available, but there were flaws in them. For example, one of the quarterbacks in the NFL dataset was a quarterback for several years but then switched positions, so his statistics will look very different from those for other quarterbacks. We also pulled the data from two entirely different datasets, so we did not have complete NFL and college data for the players we used as data points. If we found one dataset documenting players’ four years in college and first few years in the NFL, our models may have looked much different.

Additionally, there is a lot more that goes into a “good” NFL player than his performance in college and measurable statistics. Take Tom Brady, for instance, who was not a standout in college, performed horribly at his Combine, and was drafted in the sixth round. Today he is one of the best quarterbacks to ever play, and most of his success could not have been predicted by his play in college. Qualitative characteristics are important to consider when trying to predict NFL success as well, and that is nearly impossible to do with a regression model and the data available on players.

Link to Github Repository: <https://github.com/LibbyCzarniak/Final-Project>

Link to Deepnote: <https://deepnote.com/project/146af6de-f45f-4b62-a819-8e5a1cb21de6>

Link to Dashboard: <https://secret-gorge-55796.herokuapp.com/>

**Appendix**

Exhibit 1

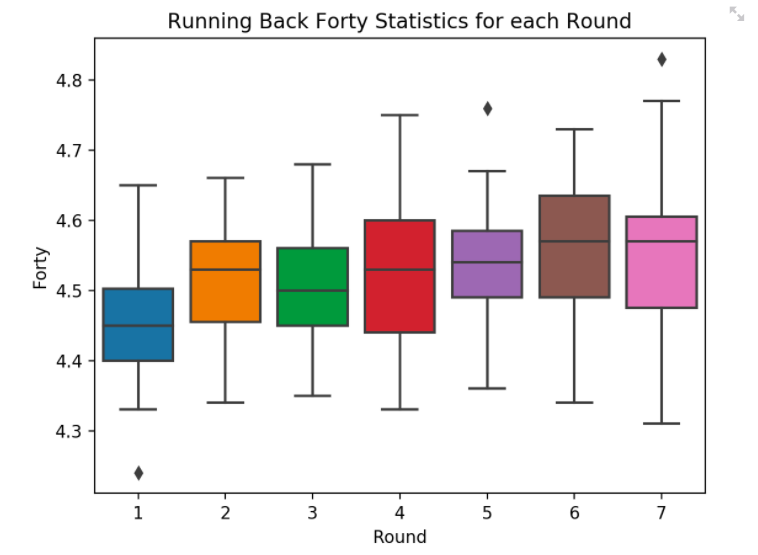
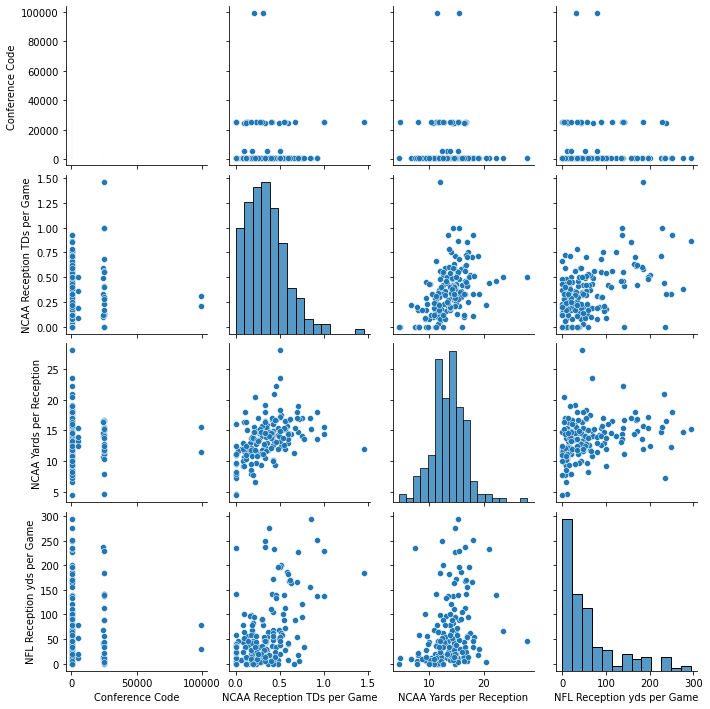


Exhibit 2

Wide Receivers and Tight Ends



**References**

<https://www.w3schools.com/python/python_ml_multiple_regression.asp>

<https://www.geeksforgeeks.org/linear-regression-python-implementation/>