NFL Attributes:

What is important?

Each year, NFL teams spend millions of dollars signing players to contracts without knowing if the player will underperform or outperform the contract. Will the player who had an amazing season last year have another one this year? Or did he just try extra hard the year before so he could get a big contract and now he intends to coast for the rest of the year? There is also a scouting event for select college players where teams record the potential players speed, strength, and agility. Are there any positions where a team just takes the player with the best result in an event? This project is designed for NFL teams that are trying to decide what makes a good NFL player. By using linear regression, I hope to find what attributes are most important for each position and then create models for each position so teams can know what players to sign and which ones to avoid.

**Data Wrangling**

Cleaning Steps

To find the data, I went to Pro Football Reference and found spreadsheets with NFL combine statistics. I also found csv files from the last four years with the Madden franchise player ratings. The data was not uniform, especially in the format of player names. To fix this problem, I replaced the commas and slashes with spaces and broke each string into a first name and a last name. I then merged the first name and last name column so every players’ name was in the same format.

Once I had all the data cleaned and together, I concatenated all four years of Madden data vertically and then removed all duplicate entries by name after the first occurrence to prevent multicollinearity problems while doing future analysis. This also created one instance of each player and was their most recent season. After doing this, I merged the Madden data with the combine data using a left merge so only the combine numbers with corresponding Madden players were merged.

I also made a separate spreadsheet that had of the data merged by player name. This data had statistics for every season played by every player in the Madden data frame over the last four years. I created this data frame to create 2, 3 and 4-year averages of each player to create a variable to account for players who were better and played longer.

Combine Ratios

Each combine event measures a different attribute, so I thought it would be interesting to see if different ratios were more strongly correlated. Does a player with a high bench press and a low vertical differ from players with a low bench press and a high vertical?

Averages

To account for players who played the most years (and would be the best players), I created a different data frame that merged all the columns horizontally by player name. I then created separate columns with a four, three, and two-year average.

Missing Values

When doing this analysis, there were two different types of missing values. The first was from the Madden data and existed because not all players played all four years. For this data, I simply replaced all of the missing data with 0’s because this would cause lower averages in the columns this data was needed for.

The second type of missing data was from players who did not participate in the combine. To replace these values, I broke down all the players into position groups and filled each missing value with the position mean for each column. I wanted to use position means instead of overall means because there is such a wide difference in the size of players on the data set.

**Other Data Sets:**

One of the things I would most like to improve for this model is to get better data. The Madden scores are a good starting point but so many of the traits are intangible still. Also, a lot of the data is not as detailed as it could be. For instance, for this dataset, there is only one category for run blocking. That category could be broken down into the components that make up the run blocking skill so more detailed data could be generated.

Another source of data that could be used would be college and professional statistics. These would help the model by showing how players deteriorate as they age.

Inferential Statistics

Defensive End:

One of the most interesting correlations is between right defensive ends and left defensive ends. It appears that weight has a correlation of 0.3233 with ‘Overall Rating’ for right ends but a correlation of essentially 0 for left ends. Right ends also need to be stronger and better at block shedding while left defensive ends need to be better at Pursuit and the mystical Awareness trait. Perhaps one side of the line is more likely to deal with the larger blockers while the other side is designed more for speed.

Linebackers

Though there are three different linebacker positions, they are not the same. Middle linebackers often drop back a few yards back into a zone coverage so it makes sense that zone coverage would be one of the positions significant correlations. Left outside linebackers seem to face more blockers based on their need for block shedding ability. Right outside linebackers on the other hand, seem to be more speed based. In football, there is usually a strong side linebacker and a weak side linebacker instead of a strict right and left linebacker so I am curious if strong side linebackers are rated as left outside linebackers and right outside linebackers are rated as weakside linebackers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute LOLB | Correlation | Attribute MLB | Correlation | Attribute ROLB | Correlation |
| Play Recognition | 0.9266 | Awareness | 0.9229 | Awareness | 0.9192 |
| Awareness | 0.9263 | Play Recognition | 0.9159 | Play Recognition | 0.9269 |
| Pursuit | 0.6945 | Pursuit | 0.7591 | Pursuit | 0.6607 |
| Tackle | 0.6277 | Tackle | 0.7222 | Tackle | 0.5607 |
| Block Shedding | 0.5756 | Block Shedding | 0.5671 |  |  |
|  |  | Zone Coverage | 0.5662 |  |  |

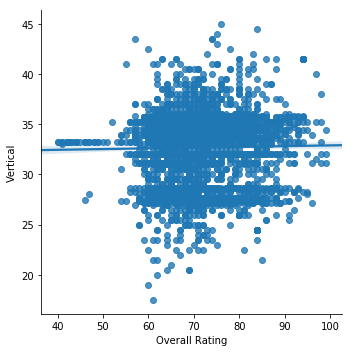
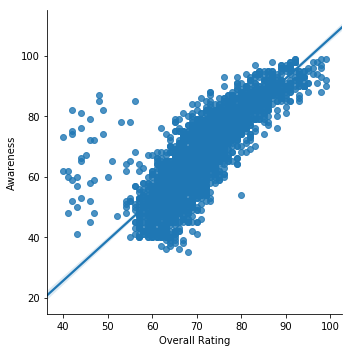
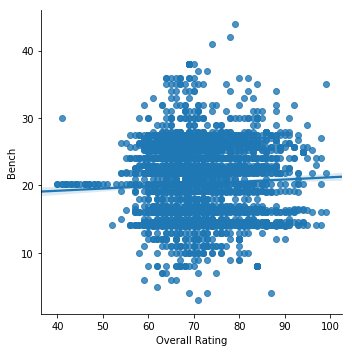
Combine Event Correlations

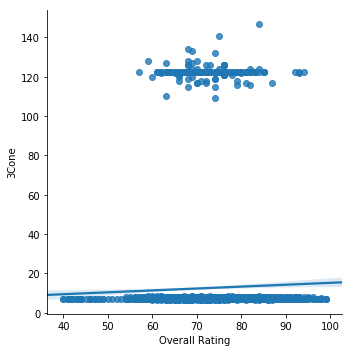
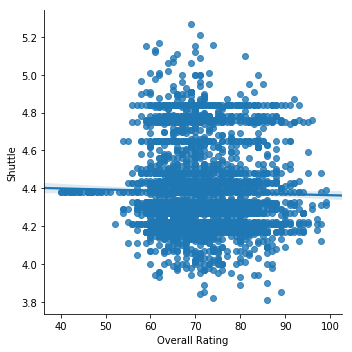
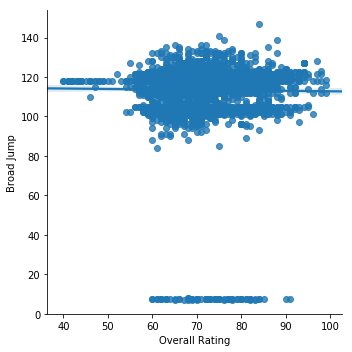
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Combine Events** | **Bench** | **Vertical** | **Broad Jump** | **3Cone** | **Shuttle** | **Height** | **Weight** |
| **Bench** | 1.0000 | -0.5411 | -0.3408 | -0.1432 | 0.6749 | -0.0114 | 0.6989 |
| **Vertical** | -0.5411 | 1.0000 | **0.4557** | 0.1848 | **-0.7854** | -0.0327 | -0.7364 |
| **Broad Jump** | -0.3408 | 0.4557 | 1.0000 | 0.1133 | -0.4107 | 0.0139 | -0.3557 |
| **3Cone** | -0.1432 | 0.1848 | 0.1133 | 1.0000 | -0.1802 | -0.0563 | -0.1849 |
| **Shuttle** | **0.6749** | -0.7854 | -0.4107 | -0.1802 | 1.0000 | 0.0177 | 0.7847 |
| **Height** | -0.0114 | -0.0327 | 0.0139 | -0.0563 | 0.0177 | 1.0000 | 0.0359 |
| **Weight** | **0.6989** | **-0.7364** | -0.3557 | -0.1849 | 0.7847 | 0.0359 | 1.0000 |

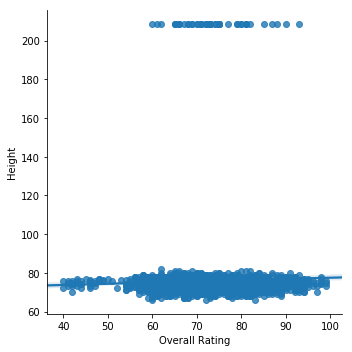
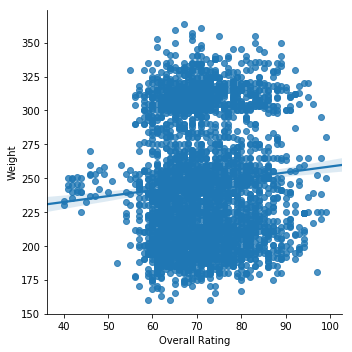
Above are the correlations for the combine events with each other. In general, the bigger the player is, the stronger they are. Conversely, the smaller a player is, the faster they are. It also looks like players who are fast can typically jump high and far as well.

**Awareness Correlation vs Combine Event Correlation**

Pictured from left to right: Bench Press, Awareness, Vertical, Broad Jump, Shuttle, 3 cones, Weight, Height







Pictured above are eight different scatter plots showing the seven different combine drills along with the most highly correlated attribute, awareness. As you can see, the physical attributes are essentially uncorrelated (weight is significantly correlated with some the offensive and defensive line positions) with Overall Rating. While there are certainly athletic minimums for the NFL (I would not be able to play no matter how high my awareness was), it seems that once a player reaches a certain minimum level of athleticism, it quickly drops in importance and other, non-physical attributes become more important.