Capstone 1 Milestone Report

What is the Problem?

The NFL's Next Gen Stats has partnered with Kaggle.com to see who can best predict how many yards an NFL player will gain after receiving a handoff. The NFL has been around since 1920, and over the last 100 years coaches, players, and fans have accumulated a treasure trove of conventional wisdom about what makes a team, player, or play successful. Recently, the NFL has begun to follow the footsteps of Major League Baseball by using sophisticated data analytics to find out where conventional wisdom is right, and where it might fall short.

There are two main use cases for a model that predicts rushing yards. The first is for the benefit of NFL teams who can take insights from the model and apply them to their own strategy to improve performance. The second is for use during the telecast or pre-game/post-game breakdowns. The NFL makes more than $6 billion annually from TV deals, and this viewing experience is enhanced by highly knowledgeable sport's casters, commentators, and analysts. If a predictive model can be used by these experts to further enhance the viewing experience, the potential value to the NFL could be in the tens of millions.

Data Wrangling

I used data from the NFL Big Data Bowl Kaggle competition. The data was imported directly from Kaggle and contained 682,154 rows and 48 columns. Each row represents a player on a given play, and because there are 11 offensive and 11 defensive players per play, every batch of 22 rows represents a single play. Thus, there were 31,007 plays to be analyzed in the data set.

Variations in spelling/misspelled words

Although the data was provided by Kaggle, there were still several cleaning steps needed before analysis. The first was with the StadiumType variable, which contained many misspellings or alternative spellings of the same word (ex. Outdoor, outdoors, outdor, Ourdoor, etc.) There were 33 unique variations which I manually recoded into 2 categories: indoor and outdoor.

A similar variable was Turf, which described the field of play. Again there were many variations of spellings, which I manually recoded into two categories: natural and artificial.

The team of the player was represented by a three letter abbreviation, however there were some inconsistencies. For example, the Arizona Cardinals were sometimes represented as ARI and other times as ARZ, so I found the few offenders and recoded those to be consistent.

The WindSpeed variable sometimes contained a single number, and other times contained the 'mph' unit at the end, so I stripped 'mph' from every row. Other rows contained two numbers (ex. 12-22), so I took the average of the two numbers in those cases.

The WindDirection variable also had variations (ex. ‘N’, ‘North’, ‘from the north’, etc.), so I converted these to be consistent acronyms (ex. ‘N’, ‘NW’, ‘SE’, etc.). Because there are 8 possible directions I assigned values between ⅛ and 8/8 starting with ‘N’ and moving clockwise (‘N’ = ⅛, ‘NE’ = 2/8, ‘E’ = ⅜, etc.).

The Weather variable had many possible variations, so I looked at the most common words and decided to make a numbered scale that ranged from -3 being most inclement to +3 being most fair. Mention of snow was given -3, rain was -2, cloudy was -1, clear was 1, sunny was 2, and indoor stadiums or closed dome stadiums were given a 3. Any weather condition that could not be placed in one of these categories was given a 0.

Converting to common units

The GameClock variable shows how much time was left in the quarter and was in the format Minutes:Seconds:Hundredths of a second. Because the data was collected once per second, the hundredths were always zero. I converted this variable to be the total number of seconds remaining in the quarter. The height of players was in the format Feet-Inches, so I converted this variable to total inches. The TimeSnap, TimeHandoff, and PlayerBirthDate variables were in the form of a datetime object, but were imported as a string object so I converted them to datetime.

Standardizing all plays to go from left to right

About half of the plays moved from left to right, while the other half moved from right to left. This is important because the X and Y positions, and the Orientation must be interpreted relative to the direction of the play. The field is 120 yards long (including the endzones) so if the play was going right to left I changed the X position to be (120-X). The field is 53 and ⅓ yards wide, so for plays moving right to left I changed the Y positions to be (53.33-Y). The orientation variable was in degrees from 1-360, so if the play was right to left I changed it to be (180-d).

Missing Values

The variables with missing values were FieldPosition, StadiumType, Temperature, Humidity, WindSpeed, and WindDirection. FieldPosition is irrelevant because that can be determined by the YardsLeft variable which has no missing values, so FieldPosition was dropped entirely. For StadiumType, I created a dictionary of every team and the type of the home stadium. For every missing row I used the HomeTeamAbbr variable to assign the StadiumType.

The Temperature, Humidity, WindSpeed and WindDirection variables were all missing for the indoor stadiums. I did some research and found that indoor football stadiums are usually set to 65 degrees F for games so I filled in all missing values with 65. I could not find the typical humidity of indoor football stadiums, but I did find that the typical humidity inside a home is about 50%. The average Humidity in the data set was 54.76 so I used mean replacement for the missing values. There is no wind in an indoor stadium, so I set the missing WindSpeed and WindDirection variables equal to 0.

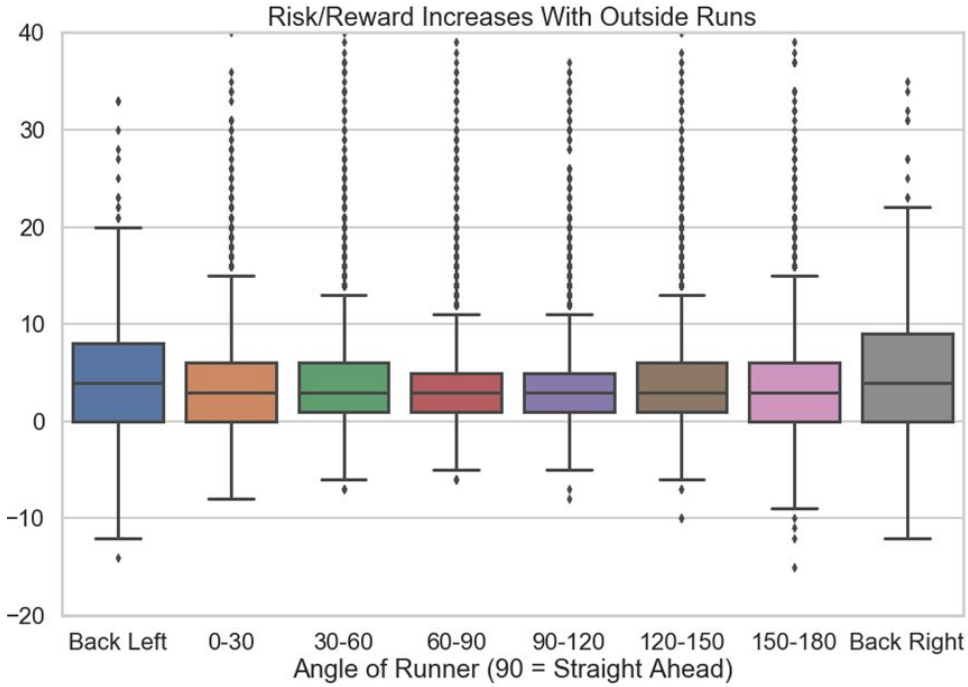
Outliers

I expect X, Y, and Orientation to be the most important predictors, and they are bound by the field of play and the possible 360 degrees of orientation, so I do not consider any of these to have outliers. I checked all of the numeric variables for outliers using a boxplot, and while some measurements were outside of 1.5x the IQR, they were all realistic values in a game of football and did not seem like they were measured incorrectly, so I did not remove any outliers from the data set.

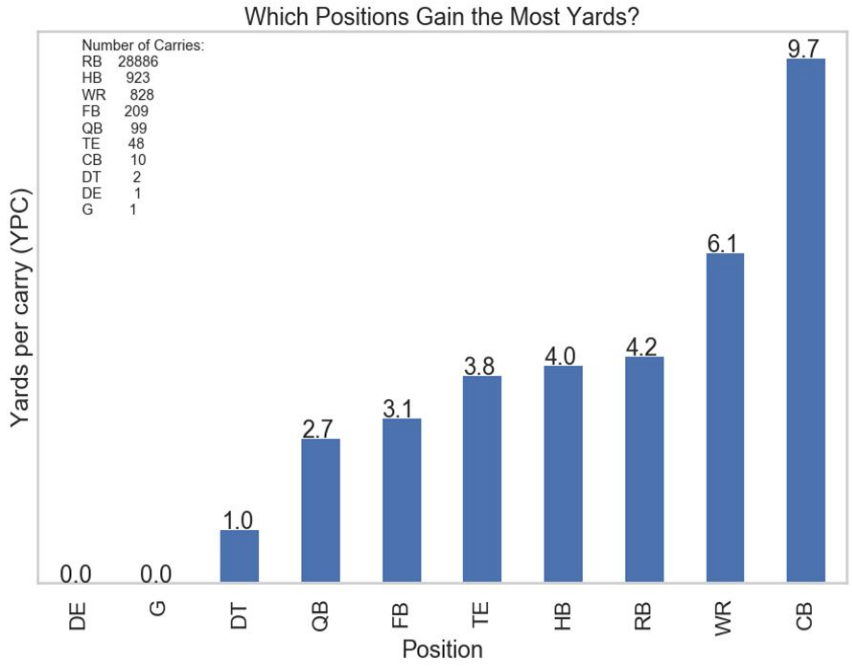
EDA Findings

Most of the exploratory data analysis involved identifying variables that are correlated with yards gained. Many of these findings agree with conventional wisdom about football. For example, teams tend to gain more yards when there are fewer defenders “in the box”. “Getting the ball in space” is a common phrase in football that refers to putting players in position to utilize their athleticism to juke and outrun defenders, and it appears that the data corroborates this idea. Runners who get the ball with more space between themselves and the nearest defender tend to gain more yards.

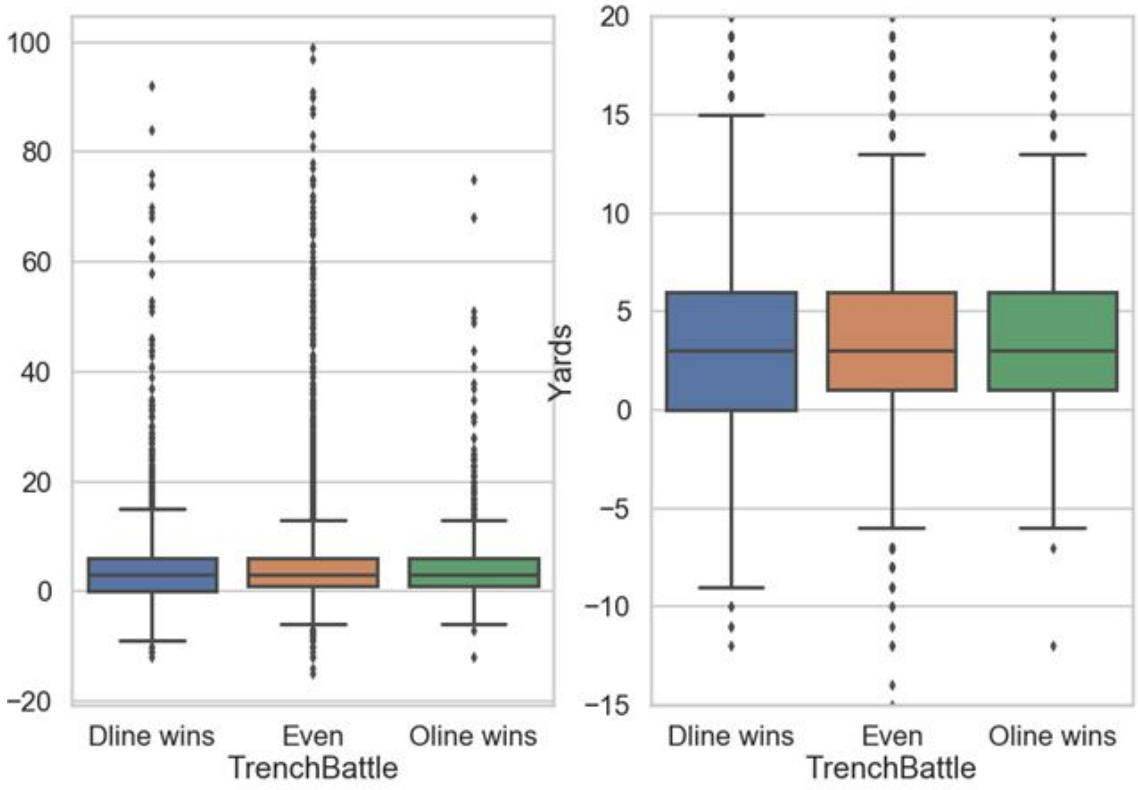
Another interesting finding was the relationship between the runners direction at the handoff and the yards gained. NFL commentators often express a preference for “North-South” runners (those that run directly towards the opponent’s end zone with little deviation left or right). The idea is that running “North-South” is safer, while dancing “East and West” has a higher risk/reward ratio. Again, the data seem to agree. The more a player is headed toward the sideline when receiving the handoff, the higher the variance in yards gained.



Running backs and half backs (a position nearly identical to running backs) carry the ball on 96% of running plays. Wide receivers make up most of the remaining 4% (828 total), but gain nearly two extra yards per carry (4.2 vs. 6.1). This is an enormous discrepancy in football; for context the highest career YPC for a running back in the history of the league is Jamaal Charles with 5.4. While this may seem to indicate that teams should hand the ball to their wide receivers more often, some of the success of wide receivers is likely due to the surprise factor from how rarely they run the ball.



A potentially surprising finding has to do with how the battle between the offensive and defensive lines affects a running play. When the ball is snapped, the offensive line tries to push the defensive line backwards to give more space for the runner to run. Conversely, the defensive line tries to push or penetrate the offensive line in order to get to the runner. To get an idea of which line was successful in that objective on a given play, I calculated the average position of all linemen relative to the line of scrimmage. Surprisingly, on plays where the average position was behind the line of scrimmage (success for the defense) the yards gained had a higher variance. The risk/reward ratio was much higher on these plays. This presents an opportunity to further investigate what separates the big gain plays from the big loss plays when the defensive line gets in the backfield.



The most direct and significant correlations found had mostly to do with the physics of the runner at the handoff. Runner's with a higher speed and acceleration at the handoff tend to gain more yards.