NFL Play by Play Prediction- Midterm Report

Jibran Gilani (jg793), Milan Shah (mrs282)

November 2019

1 Data Description

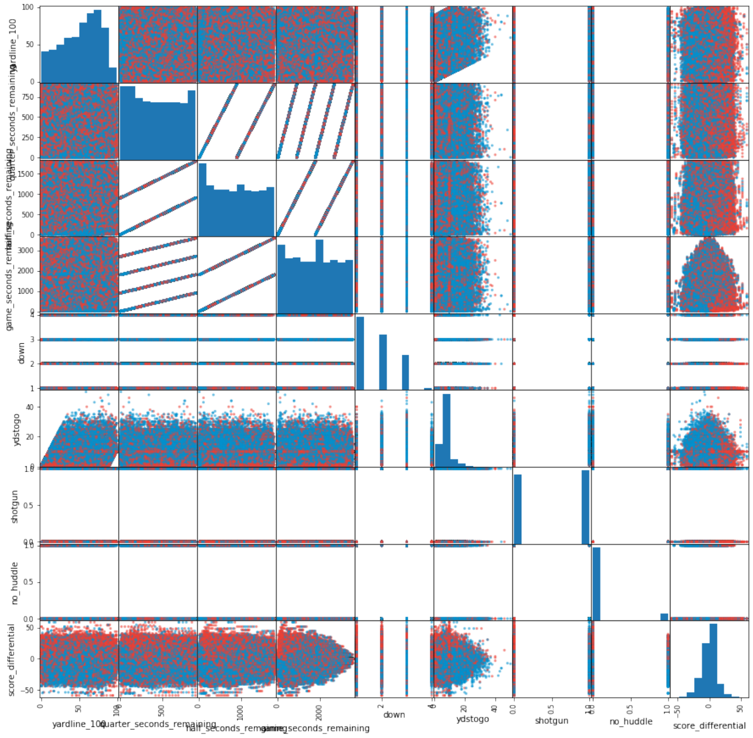
After analyzing the data, there appears to be a division into three distinct categories- game factors, play results, and probabilities. The game factors include information such as time left, score, down and distance etc. and is what we will use for our predictions. We are not as concerned with the result of the play since we only what to predict the type of play beforehand. We are unsure how exactly the probabilities are computed, but perhaps they can be used in later analysis. Examples include touchdown probability, which should be the likelihood that a team scores a touchdown on the given play. We are using XYZ columns for our final dataset, and dropping all rows that are not a pass or run play (other plays include special teams plays, such as field goals and punts, and also specialized plays including QB kneel-downs and spikes). In the future, we may want to have multi-class classification, but for now we are focusing only on run and pass plays, by far the two most common plays. We have over 300,000 plays and have narrowed down the features to the 9 most important ones that sufficiently portray the game situation.

2 Features

We began by choosing features that we categorized as game factors. We did not want to use probabilities just yet as they are not as readily available to NFL coaches and staff as game factors are. We began by treating each play as its own entity. This meant that we only considered features if they would be know by simply looking at the field. This included quarterback position, yard line, down, yards to go, time remaining, and scores among a few other features. In total we used 16 features to train our Random Forest model. Before deciding which columns to use, we created a scatter matrix to graph all the features against one another along with their corresponding histograms of values. We plotted the scatter plots using the play label as the color of the data point. This helped us determine which features and which pair of features show relevant information. We looked for pairs of features where there was a distinct split between red and blue points in order to determine the features that can be best used to predict the best play.

We used these game factor features as our base for X, but we were still missing many more features. Since football is a game dependent on previous actions, we decided we could view this problem similarly to a time series problem where features can include results from previous states in the game. These can include statistics on previous plays run in the game and even what play a team choose to run when faced with similar game factors previously in the game.

We have implemented game factor features into our model, but have not yet considered time series data, but we plan on adding it soon once we pre-process the data to include the information we find relevant.



3 Cleaning

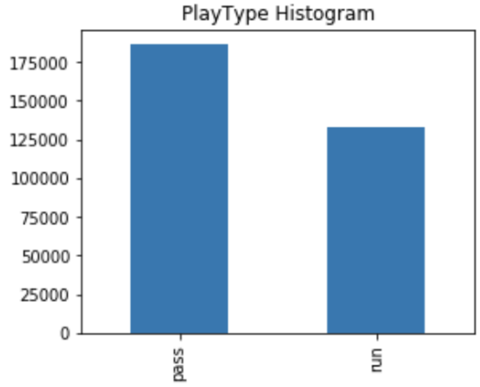
The dataset was mostly clean except several columns had NaN values, which we were able to drop given the size of our dataset. Most of the columns in the data did not have outliers except for "Air Yards" which are the length that the ball travels in air during a pass play. There we found several examples with large negative values, which seem to be typos. However, we did not use this column in our dataset since it is a feature describing the result of the play. After looking at descriptive statistics (min, max, median, mean etc.) for the other features, none appeared to have outliers (downs were from 1 to 4, yard line from 0 to 99 and so on, meaning that the features made sense in football game-play terms).

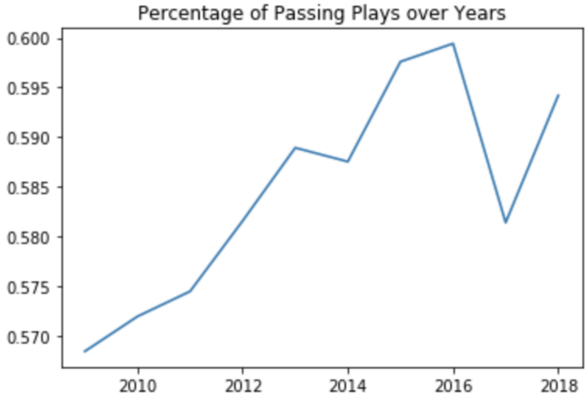
4 Train, Validate, Test Split

Our data does have time-series elements given that time within the game and season do matter, however we found little evidence of trends across years. Currently, we are training on the first 7 years of our data and validating on the last, but might transition to "leave-one-out" where we leave one year out and train on the other years for each year. For further testing, it could be interesting to see this used in real life as games from the 2019 season are underway. In reality, teams will likely game plan for specific offensive coordinators / teams, so we can also filter our data to only have plays that were called by a specific OC and fit a model to those tendencies. However, for now, we are looking for overall trends across all teams.

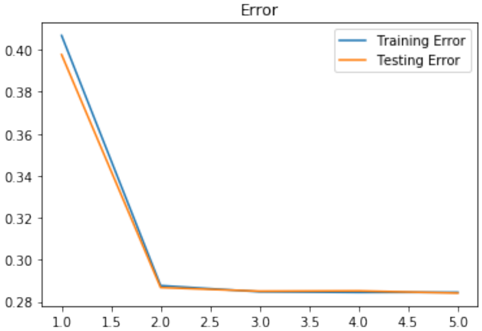
5 Model and Errors in Fitting

The first model we used was a Random Forest model, and to avoid overfitting we limited the max depth to 5. Size of the dataset should not be an issue since we have every play for the last several years, and if we introduce anymore we may find yearly trends. Over the past several decades, teams have shifted away from a run-first approach into a pass-first approach, but our dataset appears to have consistent ratios of runs to passes (around 58% pass).





We decided to train several Random Forest models each with a different depth to analyze at what point the model stops improving. Below we graphed training and testing accuracy by tree depth of the random forest, and as seen, both accuracies decrease to 28% by a tree depth of 2-3 and cease to improve from there. This is a fairly low tree-depth and we want to make sure it isn't too high to prevent overfitting. The features we choose worked for certain circumstances, but not for many others.



We still plan on training more linear models like SVM to compare and see which could possibly work best with the current features we are using and the features we plan on using. One major benefit to the random forest was that it is quick to fit and train, while SVMs take a while.