NFL Matchup Analysis

This project was intended to solve the problem of predicting the winner of an NFL football game before the game starts. In other words, based on the matchup of two teams I wanted to know which one has the advantage in the game.

This model would be valuable to a number of groups. In a very direct way anyone who wants to predict the outcome of a game could use this model to either increase the confidence of their decision or help them to decide when they are unsure. This would apply to betting platforms, people using a betting platform, and sports analysts who predict games as part of their job. A second and more notable group that would benefit from the model are people who are trying to strengthen and manage a team of their own such as coaches, general managers, and owners. They wouldn't care directly about the predictions of the model but more so how the model is configured. By seeing what the model values in a successful team, managers will know where to invest their resources. By understanding the relationships between the features of the model, teams can develop plans for taking advantage of an opponent's weaknesses or prepare by patching up weaknesses of their own.

My approach to this problem utilized an NFL API in order to gather the data in a way that I can control. I chose a large collection of features focusing on covering as many aspects of the game as possible while also leveraging my domain knowledge to select valuable statistics. The API was able to generate a broad range of information and statistics so before I began I had to decide the features I wanted and formalized my approach.

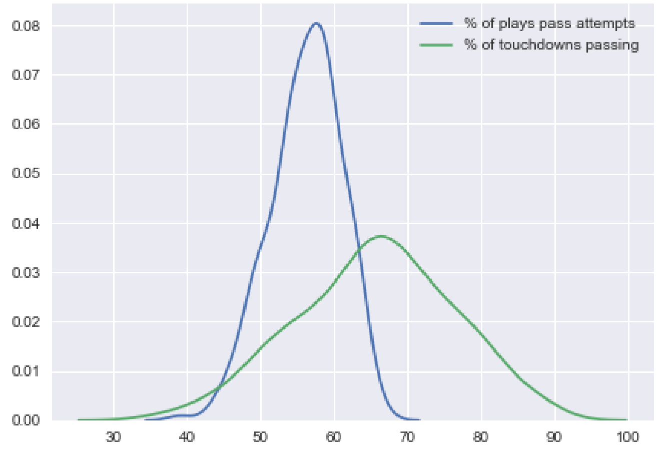
First, I decided that I wanted features that could be easily controlled by coaches and organizations making the predictive model useful as a tool for insight. By using controllable features coaches could adjust their approach to what the model values in order to better their team's performance. For this reason, I stayed away from features such as team record or previous matchup outcomes and instead tried to generate features representing team composition such as a strong running game or a good pass defense.

I began by collecting the raw data from the API, which was returned in JSON format and had to be parsed and converted into pandas dataframes. Games and team statistics were simple and involved basic manipulation of the columns and some grouping before they were saved to CSV files. However, individual player data was stored beneath multiple layers of dictionary keys and needed a team to be specified for each request. I created a function to parse the players data for each team and aggregate them into the relevant dataframe - passing, receiving, rushing, or defense where it could then be saved and manipulated later. Next, I had to merge the player data with the game data to correlate the year with the player statistics and then used various grouping, merging, and column manipulations to get the fields I desired. Once the game outcomes, team, and player data were collected the datasets were merged together to form a final dataset. Each row represents a game between 2009 and 2017 where the scores and yearly statistics of the home and away teams are the columns. The model will use the yearly team data for each team to try and predict the outcome of a given matchup.

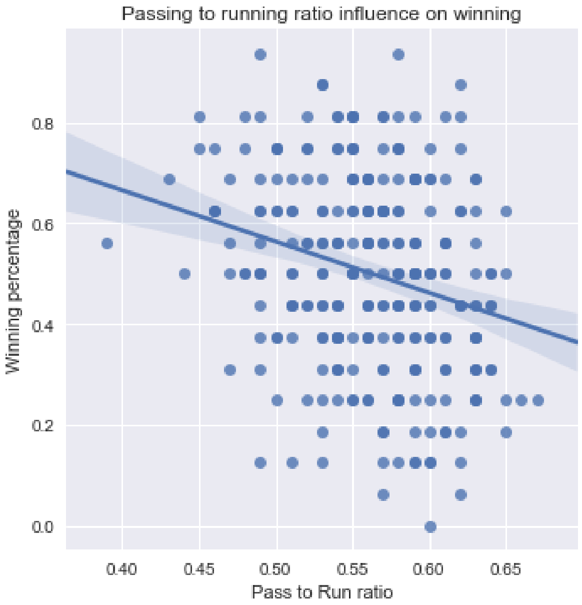
After some validation queries I could see that the API was missing player statistics for 5 teams since 2009. Games which involved one of these teams were excluded, however, this does not seem to be an issue as there are still over 2,000 matchups for training the model. Because the data had been collected in a custom way no extra cleaning steps were needed.

Once the data was collected I began to explore it in ways I thought interesting and useful from the perspective of the problem. I then tested the insights I found using statistical tools to get an understanding of its significance.

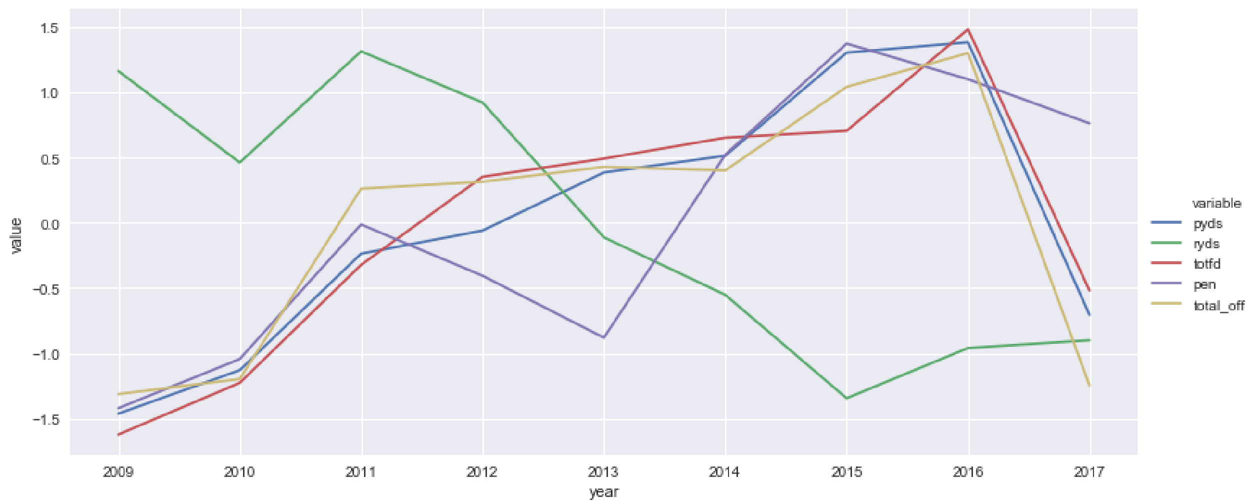
The first topic I analyzed was the team play composition. We looked at the distribution of pass to run play ratios as well as pass to run ratios for scoring, quarterbacks, and running backs. For play calls, passing was firmly the dominate play type accounting for about 57% of plays and even higher for plays that scored. I hypothesized that overall passing is a more effective method of getting yards and scoring so teams that are good at passing pass often and are able to win while teams that are bad at passing are forced to rely on a run game, an overall less effective strategy.



To test this claim I generated a scatter plot showing the correlation between passing and winning and believed that high passing percentage teams would win more. However, the opposite turned out to be true. In fact, the correlation was negative and statistically significant. I then changed the assumptions to the following in order to fit the data. Passing plays are more common because there are more situations where running isn't viable. When a team is winning it is advantageous for them to run the ball at the end of the game, and because of this it may be that winning causes teams to have a higher run percentage instead of the other way around.

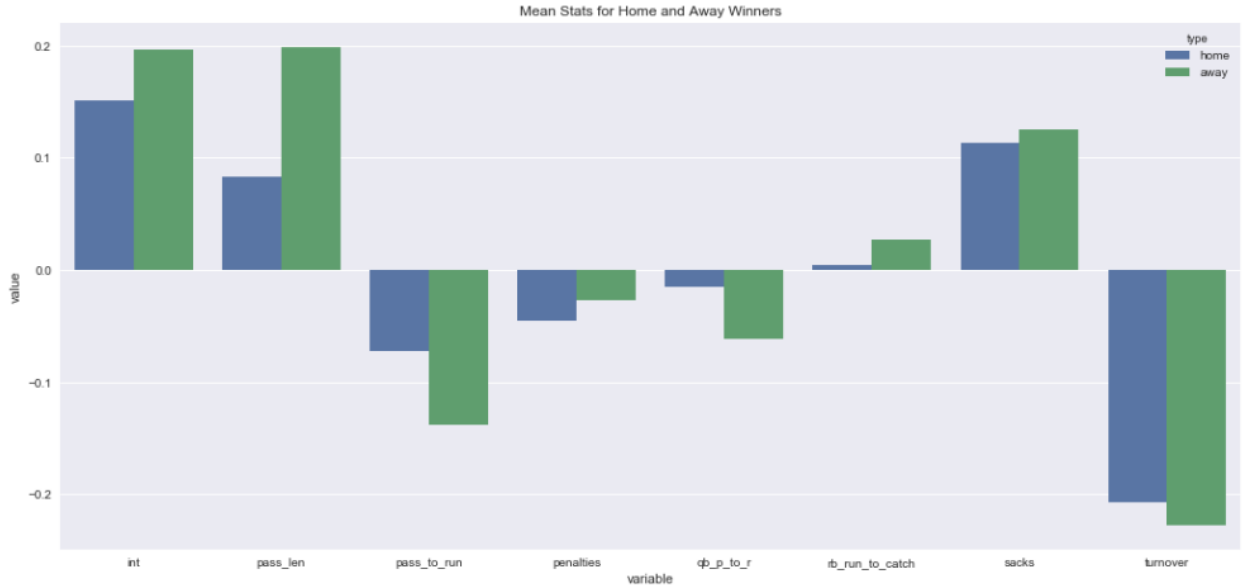


Next, I decided to look at the normalized trends of league wide statistics over the past eight years in order to gain insight how the league has changed.



The first observation is the strong correlation between passing yards, first downs, and total offensive yards. As shown earlier passing is the dominant play call so as the yardage for passing fluctuates first downs and total yardage should change with while being relatively independent of rushing yardage. Another thing to note is the increase of penalties (purple). I believe this is due to the recent rule changes the NFL implemented to prevent head injuries after being pressured from the media and its fans.

I then wanted to look at the mean of normalized statistics for teams that won games and compare both home and away. That is, for teams that won what were there average statistics compared to all teams. By doing this I could see what traits are key to a winning team and if winning takes a different type of team depending on whether the team is home or away.



For this chart it appears that turnovers are a critical part of winning. The most extreme statistics for teams that won are a high amount of intercepting your opponent (a positive turnover) and a very low number of turnovers given away. Also, low passing percentages are seen for winning teams which align with our previous observations.

One interesting observation of the above chart is that teams that win while away have similar statistics to teams that win while at home except they are more extreme. Because of this I proposed that winning while away is in general harder and therefor the statistics of teams that won while away were more pronounced in these positive ways than the teams winning at home had to be. In order to further investigate this claim, I calculated the difference in winning percentage of home vs away teams. Home teams are nearly 15% more likely to win than an away team. By conducting a hypothesis test I determined the probability of this occurring due to chance is less than 1%, showing that home field advantage is very likely a real factor in competing teams.