NFL Matchup Analysis

1. Context

This project is intended as a solution for the problem of predicting the winner of an NFL football game prior to the start of the game. In other words, based on the matchup of two teams a model is designed to predict the winner.

This model would be valuable to a number of groups. In a very direct way anyone wanting to predict the outcome of a game could use this model to aid their decision-making process. This applies to gambling platforms, users of these platform, and sports analysts/analyst groups where predicting games is important for their job. A second and more notable group that would benefit from a strong model is anyone who is trying to strengthen and manage a team of their own such as coaches, general managers, and owners. They would care less about the predictions of the model but more so how the model is configured. By knowing what the model values in considering a team successful, managers can learn 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. In this way the model could be used as tool for analysis.

2. Data Collection

An NFL data API was utilized in order to gather the data in a controllable way. Features were designed and collected to measure many aspects of the game in a broad way while also attempting to leverage domain knowledge to create features highly correlated with winning percentage and minimize noise.

Choosing features that can be influenced by coaches and owners was also prioritized. This again will help to optimize the value of model by making it useful to a wider audience. Because of this point high level features, such as 'Total Yards', were decided against in favor of more specific features based on, for example, individual positions and play-call style. Overall, over forty features, twenty per team, we designed for collection.

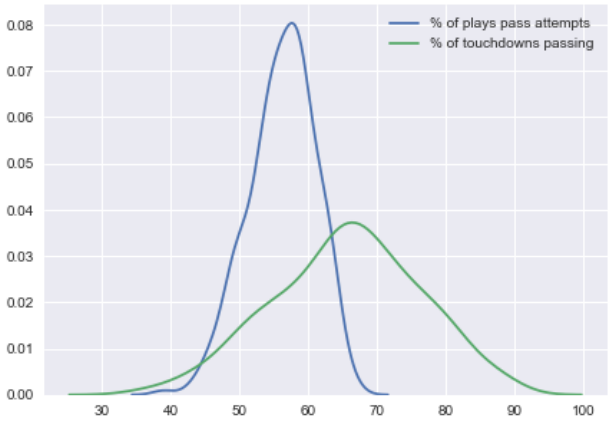
Manipulating the raw API data into the desired features posed challenges but leveraging the Pandas framework kept the process relatively efficient. Each row represents a game between 2009 and 2017 where the scores and yearly statistics of the home and away teams are the columns. The model uses the yearly team data and tries to predict the outcome by comparing the same set of statistics of the home and away teams.

After some validation queries it was discovered that the API was missing data for five team seasons since 2009, all games involving these teams were excluded, however, over 2,000 rows still exist for training and testing of the model. Because this data was custom collected no other data cleaning steps were needed.

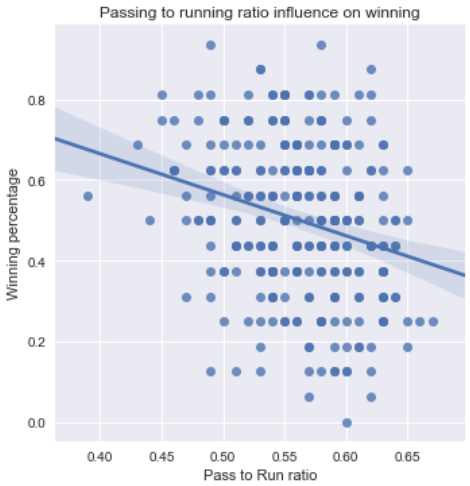
3. Exploratory/Statistical Analysis

Before building an automated model, the data was explored for trends and statistical tests were run to help generate further insights about our problem.

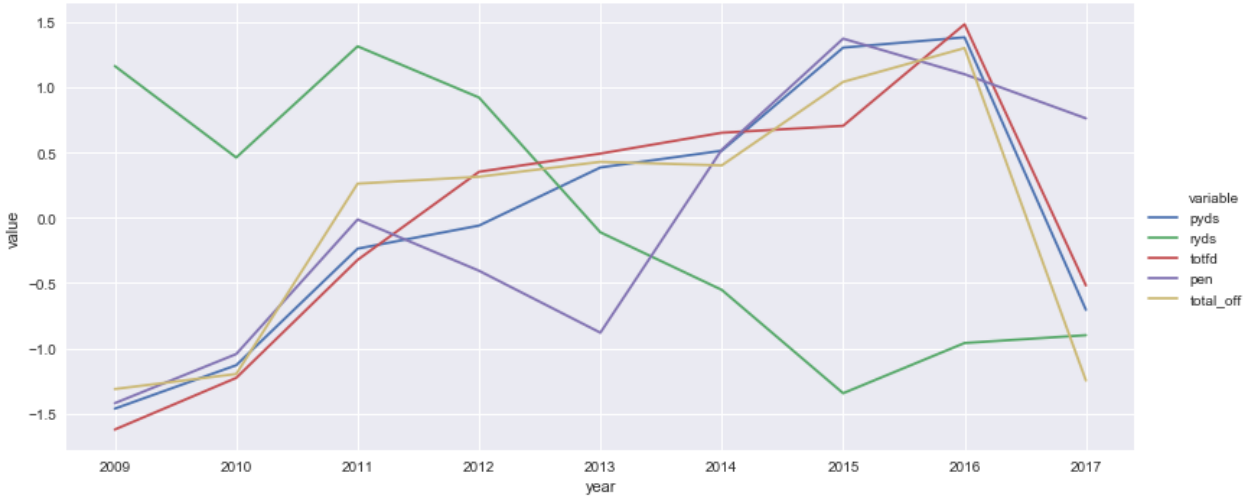
The first topic analyzed was the team play composition. We looked at the distribution of pass to run play ratios as well as pass to run ratios for scoring, quarterbacks, and running backs. For play calls, passing was the dominate play type accounting for about 57% of plays and even higher for scoring plays. One hypothesis to explain this data is the following -- overall passing is a more effective method of gaining yards and scoring so efficient passing teams are pass often and win more often while teams that are struggle at this are forced to rely on a run game, an overall less effective strategy.



To test this claim a scatter plot showing the correlation between passing percentage and winning was generated but despite the original findings the correlation was negative and statistically significant. This shows that higher passing percentages actually correlate with higher losing percentages. A second hypothesis was generated to explain both results and is as follows. Passing plays are more common because there are more situations where running isn't a viable option. When a team is winning, however, 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 aggregated statistics for the league were normalized and plotted over an eight year span in order to view league wide trends.

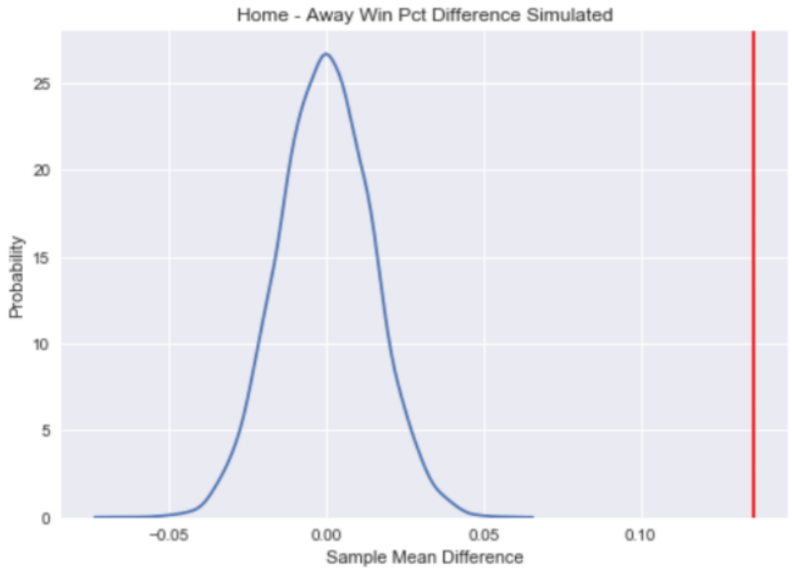


The first observation is the strong correlation between passing yards, first downs, and total offensive yards. As shown earlier passing is the dominant play call so as the yardage for passing fluctuates first downs and total yardage should be heavily influenced while at the same time being relatively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

Then the mean of normalized statistics for teams that won games were aggregated and compared between home and away teams. That is, for teams that won what were there average statistics relative to all teams. This was done for insight into the features important to a successful team.

Viewing this chart it appears that turnovers are a critical part of winning. The most extreme statistics for teams that won are a high number of interceptions (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 that the features are more extreme. Because of this it is proposed that winning while away is in general harder and therefor the features 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, the difference in winning percentage of home vs away teams were calculated. Home teams are nearly 15% more likely to win than an away team. By conducting a hypothesis test it was that 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 NFL teams. The redline below shows this 15% difference while the distribution shows the likelihood of the differences given that home vs away has no impact on winning percentage.

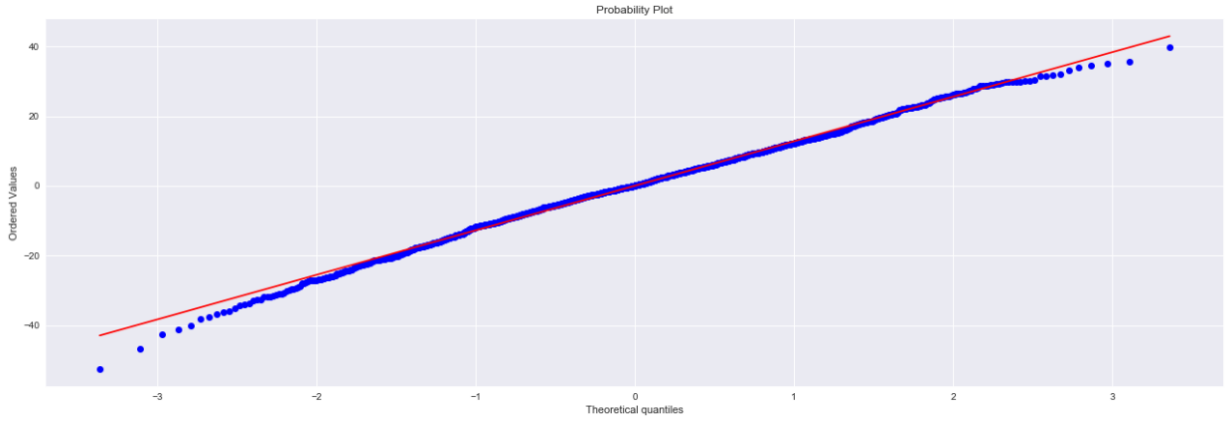


4. Machine Learning Models

The construction of the predictive model began by splitting the data into training and testing sets. The training data would be used for determining and tuning a model before using the testing set to get an unbiased measure of accuracy. The data was scaled before the models were used. This allowed for feature importance comparisons and also benefits performance for the SVM model which was used in this project. Each model had hyperparameters tuned using cross validation and incorporated some form of regularization to balance over and underfitting of the data. Within each model the feature set was experimented with and tuned for performance as well. Four models were used – a linear regression, logistic regression, SVM, and Random Forest. The analysis leveraged the Scikit-Learn library for efficient and robust implementations of these practices and algorithms.

4.1 Linear Regression

This project is attempting to solve a classification problem, however, a linear model was applied by training it to predict the score differential of the two teams. By then taking the sign of the predicted score differential the linear regression output can be converted to a classification output. It is then tuned by measuring the accuracy of this classification output to derive the best linear model. A cross validation accuracy of 69.2% was obtained making it relatively successful. Though, for a linear regression model it is assumed that the predicted values are normally distributed with respect to the correct value. By plotting a QQ chart it can be seen that this assumption does not hold and therefore weakened its validity. The distribution was asymmetrical, occasionally under predicting some scoring differences.



The coefficients of the features were examined and compared to understand their importance relative to the model. It was discovered that the two most important features were measurements from the away team, pass to run ratio and pass yards a somewhat contradictive combination. The next three most important features were turnover, pass to run ratio and passing touchdowns from the home team. Another interesting observation is that turnovers of the home team were far more important or a factor than turnovers from away teams. Feature important charts can be seen in the project code if desired.

4.2 Classification Models

The logistic regression model assumptions line up perfectly with this problem. It predicts a Bernoulli distribution given the features of the data. In other words, the probability of victory for the home team is the output after receiving a matchup as input. This algorithm also allows for visibility into feature importance by analyzing coefficients making it very convenient for analyzing relationship between features, a high priority of the project. Cross validation performance was quite high achieving 70.28% accuracy. As with the linear model, feature coefficients we extracted and examined for further insights. Many of most important features of this model matched the linear model. The one big difference was that rushing yards and rushing touchdowns was very important to this model but did not show up in the linear version.

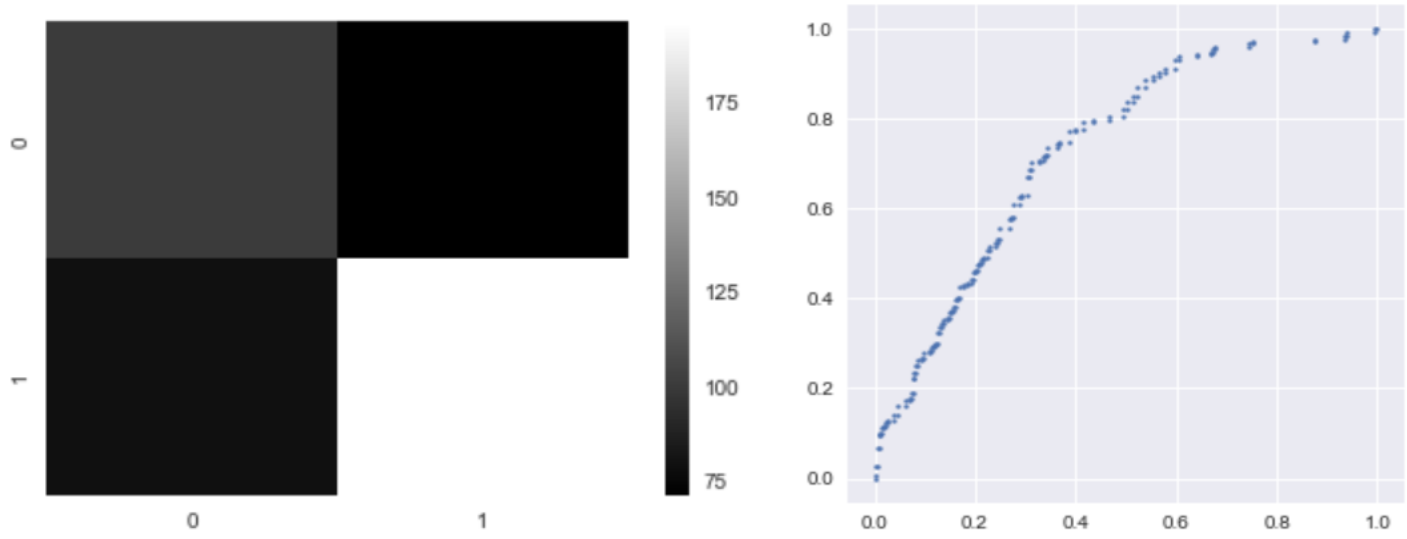
Next an SVM model was tuned. The SVM algorithm is discriminative and does not have any assumptions that need to be validated. A major drawback of this algorithm is that the feature importance is not easily visible once the algorithm has been fit to the training data. Although the model obtained a cross validation accuracy of 70.34% making it the highest performing model. Because of the nature of the algorithm feature relationships were not able to be examined at level desired by this project

Finally, a Random Forest was tried, another discriminative and incredibly flexible algorithm. Unfortunately, its performance was the worst of all the models at 63%. The reason for this is unclear and would be an interesting area of research. Although it has the benefit of feature importance visibility its poor performance kept it from being an option as the final model. Because of its poor performance a Gradient Boosted version of the Random Forest algorithm was tuned and was able to achieve a cross validation performance of 67%. Unlike the logistic and linear models, the random forest uses an algorithm which weighs the occurrences of the features within the individual decision trees to calculate an importance measure for the features

of the data. Again, similar features were important in this algorithm as were in the previous models, but turnovers for the home team was seen as a the most important feature by a significant amount. Also interceptions (a specific kind of turnover) was seen in the random forest but not seen as an important feature in the previous models.

4.3 Model Evaluation

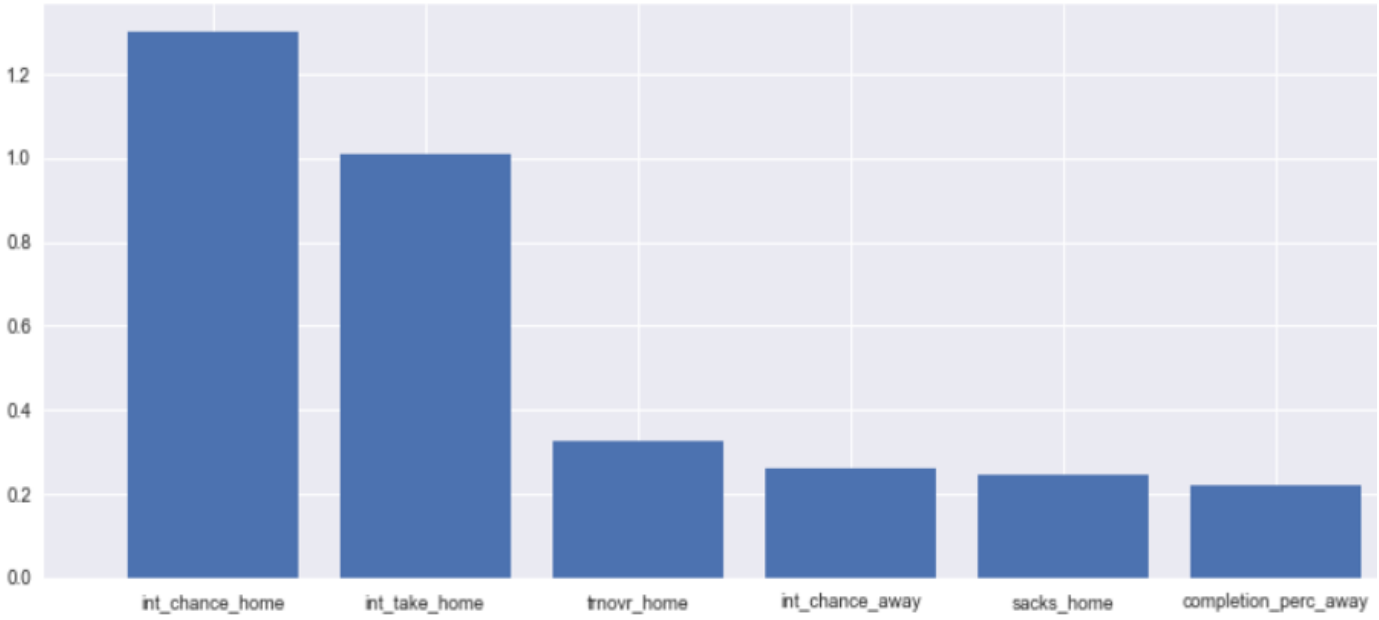
The choice for the final model, based upon performance, was between logistic and SVM which both performed very well on the training data, but because of the ability to view the importance of features after training logistic regression was chosen as the final model to be tested and used for further analysis. Upon testing the final model achieved an accuracy of 66.36%, on par with the most accurate (and some would say luckiest) analysts each season. Overall the model behaves normally, below are the confusion matrix and ROC curve visuals. We can see that the confusion matrix has balanced mis-classification with correctly identifying home wins being the most common outcome, as explained by the home field advantage statistical analysis done earlier in the project.



5. Feature Analysis

Based on our final model the most important factors for a winning team are a low pass to run play call ratio, low number of turnovers, high completion percentage (of passes by the Quarterback), high sack count, and high interception count. Interestingly the top two features were noticed during the exploratory stage of the project. Further analysis of these features and their relationships is another possible route of research. For example, discovering things that correlate or even cause changes with these ‘important features’ could be valuable information for how to improve team winning percentage. Below the most important features of the Logistic Regression model are graphed where the value of each feature is the coefficient within the logistic model, these values can also be mapped to probabilities if necessary, for an analysis.

Note the feature ‘int_chance_away’ was a combination of pass to run ratio and interception count and is by far the most important feature.



6. Improvements/Further Research Areas

A very notable property from this problem was the amount of noise within the dataset. Models were optimized by reducing the feature counts and badly overfit when the features were made into polynomials. This also limited one of the benefits of the SVM algorithm by preventing the Kernels from improving performance. Because of this, limiting the noise through different feature engineering techniques or creative feature measuring methods are key to unlocking increased accuracy in a model. One way of doing this may be incorporating statistics per players. Hierarchical models may also be a beneficial area of research as many features correlate and interact with each other. Finally, the use of time series features may also help improve performance by incorporating patterns related to team mentality and momentum.