ISyE 6740 – Spring 2021

Final Report

Team Member Names: Jacob Mackowski

Project Title: NFL Fantasy Football Initial Analytics

# Problem Statement

The National Football League (NFL) is an American sporting league that plays Football games every Thursday through Monday in from September through February. Each team has a total 53 players that are able to play on defense or offense, where the goal is to move a ball from one side of the field to the other. In a turn-based fashion, each team has up to four tries to move the ball at least 10 yards, or into the goal. If successful, their turn with the ball is extended, if they score or are unsuccessful to move 10 yards then the other team has an opportunity to try.

The analytics and sporting community have created a game that uses the players in the NFL as statistical pawns in a game called fantasy football. The goal of Fantasy football is to have the players on your team that move the ball the most down the field with a bonus if they get into the goal. At the beginning of the season there is a draft where each team needs to select the players that they believe will garnish the most points through the season. Then each week you select a lineup of your players to put against a lineup of another team’s players to see which team got the most points.

There are multiple books, podcasts, and TV shows devoted to advice and instruction for Fantasy football but most of them use “hunches” or “experiences” to give anecdotal advice. Some others also talk about analytics but there is no real statistics behind the information. There have been multiple models built on this topic already, in different ways, by a multitude of people, but each one is extremely proprietary due to a good model being able to be used to bet on sports.

The goal of this paper is to start another model that can improve reader’s wins and losses in the game of fantasy football. I will start in trying to determine what conditions lead to NFL teams having the most opportunities to move the ball. The more opportunities that the team must move the ball, should equate to the team having more points for fantasy football.

# Data Source

The data for Fantasy Football and NFL is very proprietary as well as a better dataset could lead to increased winnings in sports betting. This paper uses three free data sources and engineers a dataset to run analysis on. Play by play data is downloaded from NFL Savant (1). Data about the coaches and the players on each NFL team is from Pro-Football-Reference (2), and the weather data for each game is from NFL weather (3).

There are two data sets that are created from these sources. One dataset for clustering, which uses the play-by-play dataset to understand if the ball was run or thrown each play and by what team. The second data set are features that may or may not affect the outcome of how many times a team runs or throws while trying to move the ball down field.

As this is the start of what could become a larger model the features that were chosen are basic amalgamations that are known before a game begins. Features were chosen to avoid correlation with the team or names of players. Correlation with a specific team would cause recency bias that would cause error when looking year to year. The atmosphere in which the play occurs could affect the outcome of the play calling, so included are temperature, if it was raining, if the offense was home or away, and if it was played inside or outside. The experience of a coach was considered categorically, (new, seasoned, old) to not have each team be associated with one coach. If the team has their top players at each position, they may be more likely to use one strategy in play calling rather then another one. The percentage of starters both on Defense and Offense were considered for each game.

Further on I discuss need for additional features to improve the accuracy of the model. The additional features added are a breakdown of the percentage of offense playing. It is broken down to answer the question more specifically about the four top positional players. Those for broken down categories: the starting running back playing, the starting wide receiver, the starting quarterback and the starting tight end. The rank compared to other players is also added for those positions, to help answer the question if the player is playing are they good enough to make a difference.

# Methodology

In the data set the paper uses for clustering there are four features: offensive throws, offensive runs, defensive runs and defensive throws. The goal of the clustering is to have label different outcomes of play calling to be used in a classification model. Clustering was done using the Kmeans package of scikit-learn in python. As stated in the introduction, this is to start to understand the data and how best to start analyzing, rather than an in-depth finalized model. In future experiments different packages could be used to classify the dataset, changing either distance measure type, or number of nearest points for clustering.

The number of clusters (k) was varied from 2 to 10. The minimum bound was selected because anything under 2 would no longer give classification, and the maximum bond was selected for size restrictions of each class, any larger and the number of data points in each class would reduce to insignificant levels. Each of these clustering groups was kept to be able to be analyzed in future steps. In the feature dataset, all the continuous variables are standardized, and the numerical categories are changed to categorical for modeling.

Chart, line chart

Description automatically generated

Figure 1: Increasing the number of K (x-axis) decreases the number of samples in each grouping and decreases the standard deviation of the overall groups.

The goal of the paper is to be able to use the model output for decisions in the game of fantasy football. The output of the model would be a number that corelates to a distribution of run plays and pass plays called by each team. With the assumption that more plays mean a higher likelihood of points, these distributions would be what a fantasy football player would use to determine who he should start.

It would then be advantageous to have more possible clusters, as shown above the more clusters that are predicted the smaller the standard deviations. This would be less uncertainty that the mean number of plays were going to be called, and more confidence in the start discussion. This would be the case if the number of games used in each cluster was high enough to be statistically relevant to prevent over fitting.

Diagram

Description automatically generated with low confidenceDiagram

Description automatically generated

Figure 2 (left): Histograms of 4 clusters possible game outcomes.

Figure 3 (right): Histograms of 8 clusters possible game outcomes.

In the first modeling experiment 8 clusters was used, as this would give use the smallest variance in each cluster if a good model could be found. The features that were used: temperature of arena, was the game played inside, was it raining, percentage of offensive starts playing, percentage of defensive players starting, was the offense playing at home, is the offensive coach new, is the defensive coach new. These were chosen to not correlate to a specific team or player.

I used 80 percent of the data as training data and the balance as testing data. The models that were chosen were a basic logistic regression, a random forest, boosted trees and Adaboost. All algorithms used the default parameters that come with the python Scikit learn packages. More complex manipulation of these were not done as it was seen that even incremental improvement that these may have had would not have improved the solution enough to change the outcome. Scoring was done by accuracy, which is total number of correctly grouped games over total number of games.

These models were chosen, for both speed and their ability to classify data. In logistic regression, lines are found that splits the data into the categories, with each feature having a slope that effects the slope of the lines. This model takes up low computational power and is done first because if this could work then no other method is needed. A random forest was selected because the grouping of data is interlapped in the 4D space, and whereas the logistic regression is a line, the decision tree could be a more complex boundary layer. It also maintains a very low computational requirement as it is just one model. The next two models adaboost and boosted regression are ensemble methods, that run multiple trees or regressions and give them a weight for each regression ran based on accuracy. These are used to get even more complex boundary layers and were chosen due to the high interconnectivity of the data. The basic models were also included as a base line to make sure a boosted model didn’t over fit the data.

The second modeling experiment was added to understand what path forward was best. There were two improvements form the first experiment. First, all models were tested at all cluster amounts to understand if the number of clusters could improve overall performance. More specific features about the offensive were used. The new features answered the questions; was the top running back playing, was the top wide receiver playing, was the top quarterback playing and was the top tight end playing.

In both cases, a boosted model was the best outcome. The boosted models are a combination of multiple basic models either a decision tree or a liner regression model. In the combinations, if they are an accurate model, they are given more weight in the final answer that the boosted model comes up with. If they are not accurate, then they are given a lower weight for the final answer. It aligns with logic that a combination of models boosted for accuracy would be an improvement on the basic models.

# Evaluation and Final Results

In modeling the factors used were designed to not have bias towards a team, or player. This would guard against bias for one team to have a tendency one year that may or may not be true the next year. If the machine learning model was successful it would be in its most general form and could be used across years and teams. Another factor is only free datasets were used and engineered, these needed multiple hours of engineering to be in a form usable to the model.

In the first experiment with the most generalized features, the model failed. The accuracy of the four different models tested on the data were between 12-15%. Where if you randomized the data set you would have a 1/8 chance of choosing correctly, which would be 12.5%. This would mean, with the amount of data that is available, and the features selected there isn’t a statistically significantly correlation. Two steps were taken to try and understand where analysis should move in the future.

The first step is to understand if more data, in each cluster or larger clusters could improve the accuracy of the models with the same features. The models were then run on the clusters of 4,6,8,10 and compared to what would happen if it was a randomly selected cluster. It was found that 4 clusters had a better overall accuracy and a better accuracy when compared to a random selection. The downside of four clusters is they have a larger variance in the distributions.

In experiment three more features were engineered. These features expanded one more level of detail into the percentage of offensive starters playing. There are four main starting positions in the offense. The running back, wide receiver, quarter back and Tight end. The new features looked at specifically if the starting player at each position was available for the game rather than overall offensive starters. The rank of those players was also added. All the players in the NFL are ranked based on ability, this ranking was used to determine how “good” a player was at each position. These new features improved the accuracy of the predictions as shown below.

Chart, line chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure 5 (left): Basic parameters ran through the four models.

Figure 6 (right): New Features and basic parameters run through the four models.

Random Forest: rf, boosted trees: gbt, Adaboost: abc, Logistic Regression: lr.

The goal of the paper was to start a new model, that would help improve a fantasy football player’s score. The paper outlines a basic model that starts from free base data and engineered it to give basic features. These features can be improved upon in the future as shown by experiment three, to continue to improve the analytics model. The features that were used had little to no effect on the play calling of a game. These are then by themselves variables that can be ignored. This is important because of the analysis in the media, there are multiple factors that are tested here that are used to try and persuade a decision. The reader now can ignore these arguments and decide based on other logic which has a higher likelihood to make an effect on the play calling in the game.

Although we met our goal the analytics are not helpful year to year. The amount of data needs to be increased from a one-year sample size to a multiyear sample. The starting and sitting needs to be more specific by player. Instead of using a player’s name or rank it would be beneficial to use their stats over their last few years to predict ability. This would remove the bias that comes with players names and make it more general so that one player could be switched out with another if they have the same stats. Once the features are engineered, the models need to be tuned to increase performance. There is a long way to go but knowing where to go next is half the battle.

# Citations:

1. *NFLsavant.com*. NFLsavant.com: Advanced NFL Statistics. (n.d.). Retrieved December 13, 2021, from http://nflsavant.com/about.php.
2. “Pro Football Statistics and History.” *Pro Football Statistics and History*, https://www.pro-football-reference.com/.
3. Nflweather.com. 2021. *NFLWeather.com*. [online] Available at: <http://www.nflweather.com/> [Accessed 13 December 2021].