**Predicting Fantasy Football Success**

**IST 707 – Data Analytics**

Thomas Brown

Christopher Fredrick

Samit Patel

Andrew Taylor

|  |  |
| --- | --- |
| Using Data Analytics Techniques to Determine NFL Fantasy Success | |
| **Introduction** | The NFL is the largest league for sports in America. In the 2021 NFL season, the average viewers per game was 17.1 million viewers. This is the second highest average viewership the NFL has had. The average amount of viewers went up a staggering 18% between the 2020 and 2021 seasons.  Along with the growth in the number of viewers comes the growth of the sports betting industry. There are two major parts of sports betting. This includes sportsbook betting and daily fantasy sports (DFS). Sportsbook betting is a type of betting where a user can bet on a specific thing to happen in a game. This falls under the category of gambling, therefore it is not legal in most states. DFS on the other hand is not considered betting and is considered a game of skill.  The premise of DFS is to build a virtual team and to track their success throughout a season. The reason this sort of game is lucrative is because a typical DFS user will join a league with other users, and they will wager money on whose team will do the best throughout the season. The reason this is not considered gambling is because the DFS player is not gambling on the success of a certain player. They are merely trying to build the best lineup based on statistics, which is why DFS is considered a game of skill.  In 2021, there were an estimated 57 million DFS users in the US and Canada. Approximately 40 million of those users played fantasy football. DFS had a market size of a whopping ~$9 billion in 2021. Combine that with the growth in viewers for NFL games and there is a recipe for success.    People spend hours thinking about which players will succeed in fantasy sports and which players will be busts. For the most part, it can be determined which players are going to succeed based on previous success. The purpose of this report is to bring to light some of the other players that may be overlooked at first glance. The other purpose of this report is to generally determine and cluster tiers of players for the NFL. |
| **Analysis** | **Data Preparation**  The data we will be using for the analysis exists in a github repository. This repository contains many data points such as historical fantasy points, team ratings, and most importantly, player ratings. With these player ratings, we can calculate fantasy points and can create modeling around it.  First thing first, we need to make a key assumption and set a constraint to make this an accurate analysis. The assumption we are working with is that we will be working with players that played at least 8 games per year. This is an important constraint, so we avoid players that were impacted by injury, or where we have incomplete data. To accommodate this assumption, we are going to filter the data frame.  We will also be creating a couple columns that will be used later. The first column being created is a column for average fantasy points per game. The next column that will be created is for a generic ID that will be used in the clustering that will be done in the data exploration step. In terms of preparing the data for analysis, this is all the prep we need to do.  **Data Exploration**  This part of the analysis will be the most involved part of the whole analysis. For this analysis, we will be grabbing the data from the 2016-2019 NFL seasons and combining it into one data frame. Once we do this, we will be creating our tiers. The tier system will be the most important part of this analysis. The tiers will place the players into different levels based on fantasy production per game. Tier 1 will be the highest and tier 4 will be the lowest.  The very first thing we will do in this analysis is take a quick look at our data. This is an important step, because we want to know what sort of statistics we will be looking at. It’s important to not make any assumptions in this step. We don’t want to see the first few rows and assume that there are specific data points in other rows.  We will be using these techniques for the wide receivers, but we also need to do this for the quarterbacks, runningbacks, and the tight ends. This is important because football is a team sport. The production of each position directly impacts the production of another position. We will be doing the creation of this tier system using k-means clustering. As we do this for each position, we will be using a plot function that we created earlier in the program. This plot function will show us the break out of fantasy points per game per player, per position. This visualization will help us see how things will roughly shake out when we plot the k-means clustering results. This plot will have visible columns in it. These columns will represent each year.  After we run the k-means clustering function for each of the positions, we will also run a plot function to see how our clusters look. This plot function will account for all the years in the analysis, so we will not see those columns that we saw in the original plot function. We will be doing the exact same process for all the other offensive skill positions (positions that can score points).  Using the tiers that we created with our clustering functions in the previous step, we will go through each year and put tiers on each of the players for each of the years. The results of this step will be used later in the models, but this is all for setup for those. We will be using the 2nd of the filter functions we created to break apart this data. The final step for each year is to create columns for the other wide receivers on the same team as the main wide receiver. This is important because the success of other wide receivers has a direct impact on the success of the main wide receiver we are researching.  We will continue to do this for the entire year for each position then move onto the next year and repeat the step for each position. Once we complete this up to 2018, we can move forward with combining these 3 years into one data frame. From here we will also create two auxiliary data frames that have all tiers for the wide receivers and another that has the point averages.  **Models and Methods**  For the models that we use in this analysis, we will be trying a few different approaches with a few different data sets. Below are the different methods of modelling we will be using:   * Naïve Bayes * SVM * KNN * Random Forest   **Analyses Goals and Parameters**  The goal for this analysis is to determine if our models can place players into accurate tiers for fantasy production. This analysis won’t tell you how many points a player is projected to get, but rather which players are likely to be in the top group of production. |
| **Results** | **Technical Analysis**  As mentioned in the section above, the first part of this analysis is to pull in some of the data (2016 statistics) and manually observe it. A sample of this data is visible in figure 1. Just by looking at this small sample, we can see that this dataset includes data for all offensive skill positions. One key piece of information that we can see from this small data sample is that different stats are skewed to certain positions. QBs are going to have more passing yards and more passing touchdowns, while RBs are going to have more rushing yards and more rushing touchdowns. This will be an important distinction in our analysis for creating our tiers.    Figure 1  The next stage in the analysis was creating 2 filtering functions and one plotting function. The results of those will be shown later on in the analysis. Once we create those functions, we will grab the 2017 and 2018 data. This data looks the same and is laid out the same as the 2016 data. We next combined all of the data into one data frame. This frame looked the exact same as what was seen in figure one but had more data points (1777 observations with 21 variables).  In the next part of the analysis we used one of the filter functions created earlier to create a smaller data frame, as seen in figure 2. This data frame sample has 4 data points in it; an ID, Player name, Team and the average Fantasy Points per game, which is the crux of the analysis    Figure 2  Once we have this, we create a plot using the plot function we created earlier. This plot can be seen in figure 3. This plot includes the “columns” that was mentioned in the previous section. These columns, as mentioned before, create a quasi-year breakdown. In these columns, we see the spread of average fantasy points per game. We can see an upward trend over these years in general, but we see a larger number of players getting higher points in the last year.    Figure 3  After creating this visualization, we ran the kmeans function. This lets us cluster our data into the 4 tiers that we want to create. In figure 4, you can see the result of the visualization of this cluster for wide receivers. The “columns” mentioned in the last figure are also visible in this plot, but we can see clear groups that these points and columns fall into.    Figure 4  After running a mutate function to get the wide receivers in the correct tiers, we use a group\_by function and summarize based on the tier values and in figure 5 we can see the tier ranges for each tier.    Figure 5  Next we went on to do the same thing with the quarterbacks. In figure 6, you can see the initial chart that we created for the wide receivers above. The qb breakout is much more sparse because there are much fewer qbs in the league.    Figure 6  After running the k-means clustering function on the quarterbacks, which can be seen in figure 7. We can see in this cluster that most of the quarter backs fall into the middle tiers, and there are very few that are in the highest and lowest groups. In figure 8 we can see the results of the tiers and the average fantasy points per game in each tier.    Figure 7    Figure 8  In figures 9, 10, and 11 we can see the same analysis results we saw for the wide receivers and the quarterbacks. In figure 9, we can see that the spread of fantasy points is much greater than the qbs and the wrs. The results of the k-means clustering showed the same thing for the other positions. There were a lot of players in tiers 2 and 3 but fewer in 1 and 4.    Figure 9    Figure 10    Figure 11  The final part of this stage of the analysis was the same analysis we did for the last 3 positions and applied it to the Tight Ends, the last skill position on offense. The results of this can be seen in figures 12, 13, and 14. The spread of these players is much less than that of the other positions and the groups seem to be clustered more tightly together.    Figure 12    Figure 13    Figure 14  In the next step, we placed each player in each year in a tier based on the results of our clustering. In figure 15, you can see the breakdown of only the 2016 season for wide receivers and see the spread for the season.    Figure 15  We will continue doing this for each year, but as mentioned in the previous section, when we complete one year worth of position filtering and discretization, we need to do the analysis for the wide receiver teammates. In figure 16, we will see the resulting data frame from this breakout of wide receivers. This is important as it will give us a better performing model because it will take all the skill players into account. We will continue this form of analysis through each year and for each position until we complete the 2018 season.    Figure 16  At this point we will combine all the 2016-2019 data that we compiled in the last couple stages. The summary results of this can be seen in figure 17. This will data frame will be used in our modeling functions. After this we will also be creating subset data frames that have tiers and point averages respectively.    Figure 17  The first model that we will be running is the Naïve Bayes model. The first iteration of this model is going to be with the Tier data frame that we created earlier. The purpose of this will be to test if our methodology of placing other players in appropriate tiers can help place the wide receivers into the right tier. The training set will be the tier data frame and the test set will be the tiers data frame without the tier of the wide receivers. In figure 18, we can see the results from the confusion matrix of this analysis. We can see with Naïve Bayes, we were able to yield a 31.25% accuracy which isn’t the highest possible value but tells us that our analysis is getting us on the right track. The upper end of our confidence interval is 46.25%    Figure 18  In figure 19, we ran the same analysis but set the method equal to cross validation (cv). We can see that this gave us a lower accuracy percentage, with an accuracy of 20.83%, and an upper limit of 34.99%. This particular method doesn’t inspire the most confidence.    Figure 19  In the next part of the analysis, we will be using the data frame we created earlier with points averages. For this analysis, we will be using SVM. In figure 20, you can see the results of this part of the analysis. Linear SVM yielded an accuracy of 38.71% with an upper confidence interval of 57.81%, which is getting close to what we want to see. So, it is a safe bet to say that basing our models off the average points may be a better approach. With polynomial SVM, as seen in figure 21, our results were even further improved with an accuracy of 51.61% and an upper confidence of 69.85%.    Figure 20    Figure 21  The next model we moved onto is the KNN model. These results can be seen in figure 22. The results from KNN yielded a lower accuracy compared to both SVM models with only 25.81% accuracy.    Figure 22  The final model we are going to try is randomForest. In figure 23, you can see the results of this model. This model’s accuracy came in at roughly 35.42%, which falls in line at the rough average of the results of our models.    Figure 23 |
| **Conclusions** | Based on the results of our analysis, I think it is safe to say that the average fantasy points analysis we did is the part of the analysis that has the most viability. The average fantasy points part of the analysis was the part that returned an accuracy of 50% when it was trained. The average user would also be able to come up with some sort of analysis like this, obviously not this involved, but a basic analysis like this.  One thing that needs to be pointed out is the breakout of each tier per team. This data point can show that this analysis is more than just an analysis about players, but also shows the viability of this analysis on picking which teams are likely to succeed based on which team has multiple top-level performers. This supplementary data can be seen in figure 24 below.    Figure 24  Looking at the next steps for this type of analysis, there are improvements that could be made. This is a great starting point, but there are other factors that can be looked at to further the analysis. Below is a list of a few internal/team pieces that could be added to the analysis:   * Coach * Injuries/Previous Injuries * Performance against defenses * Play formations/scheme   There are also some other external factors that could be taken into account for the next steps of this sort of analysis. These factors are things that can’t be controlled by the player but are factors that show how a player performs under certain conditions. Some of these factors are listed below:   * Performance in different temperatures * Performance in different stadium settings * Performance of referee team   As we take these results into account, it's important to note to there is one factor that cannot be taken lightly. That factor is the human factor. All these players have created careers around constituency, but it is worth noting that they are just humans. Things happen in the lives of these people that can change the week-to-week outcome. There will always be some percentage of uncertainty.  With all of that in mind, based on the results of this analysis, it is possible to successfully create groups/tiers for these players and predict which players are going to succeed in the following years. Even the fundamental part of the analysis shows that there is decent confidence in this style of analysis to work. In a subject matter that is this lucrative and has so much visibility tied to it in the sports world, this sort of competitive advantage could lead to a lot of potential money. |