**NFL Fantatsy Football Team Selection**

Principles of Data Mining

Professor Thomas Kinsman

McKay Clawson, Patrick Gleason, Kristopher Hopkins

Table of Contents

Introduction: 1

Background: 2

Data Collection: 4

Descriptions and Relevance: 5

Solution: 6

Results: 8

Conclusion: 10

# Introduction:

In the late 1990’s online NFL fantasy football was born. Since then the game has become one of the most popular friendly competitions between football fans. Co-workers, family members and casual friends gather year after year to draft a team that they hope will finally carry them to this years championship game. Fantasy football is so popular that many people draft multiple teams, pay large league entry fees, and religiously check statistics. FX has even made a very successful TV show around the concept of friends competing in a fantasy football league. It would appear that there is no end in sight for the fantasy craze that has captivated America, especially considering the fact that NFL is enjoying an all time high in popularity and television ratings. More and more we are seeing the introduction of advanced statistics to draft strategies. Looking at complicated metrics like QBR (Quarter-Back-Rating), YAC (Yards-After-Catch), RAC (Rush-After-Contact) and many more are quickly becoming the modus operandi for serious players. With all of the advanced data available, data mining should be able to help players take the step to the next level. One of the hardest parts of creating a competitive team is the actual process of drafting the team. Year after year we see players get drafted in early rounds of drafts across the country only to see them have lack luster years. We think that using data science we can help alleviate the pain of drafting players that have mediocre years. The basic idea is to use a number of different clustering and classification techniques to make a simple decision: given a player name and a round number, should you draft that player.

# Background:

Fantasy football was officially created in 1962 when the first rulebook was publically published. For the next 3 decades, the game enjoyed moderate success among the football faithful. One of the biggest reasons fantasy football never reached the popularity levels the game enjoys today was how difficult it was to play the game. The draft was a manual process, where people would sit around a table and take turns selecting players. From then on players would lock in their lineups before the first game by submitting the players they thought would do best to the league commissioner. Then on Tuesday morning everyone would look at the sports section of the newspaper, find all of his or her players and look at the stats to determine how many points they had earned that week. The advent of the Internet changed the game as we know it. In 1997 CBS Sports launched the first online client and fantasy football was born. Using different websites such as ESPN, Yahoo!, or CBS Sports makes the game much more fun because it removes all of the manually intensive labor from the game. The website will track the stats and scores, manage teams, adjudicate trades, and create playoff brackets. The game has evolved so much that there are many different rulebooks and types of leagues. We are focusing on a standard league where you are allowed to have three types of offensive players: passers (Quarterbacks), receivers (Tight Ends and Wide Receivers) and rushers (Running Backs). Quarterbacks are the most valuable players on the team by the nature of the position. They score 4 points for each touchdown pass, 1 point for every 25 yards they throw, and lose points for various missteps (fumbles, interceptions, etc.…). Receivers score 6 points for every touchdown they catch and 1 point for every 10 yards they catch. Rushers are very similar, 6 points for a touch down and 1 point for every 10 yards they run for. With the understanding of how the game is scored, the draft is very simple. There are 10 rounds and each team gets to pick one player a round. The players you pick on draft day will make up the majority of your team for the rest of the season (you can do things midseason to improve your team, such as trading, but for simplicity we are ignoring these possibilities). It is not an understatement to say the draft is the single most important day of the season. With that understanding it is easy to see why we want to accurately determine the round players should be drafted.

# Data Collection:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rk | Player | Tm | Age | G | GS | QBrec | Cmp | Att | Cmp% |
| 1 | Peyton Manning\*+ | DEN | 37 | 16 | 16 | 13-3-0 | 450 | 659 | 68.30% |
| 2 | Matt Ryan | ATL | 28 | 16 | 16 | 4/12/00 | 439 | 651 | 67.40% |
| 3 | Drew Brees\* | NOR | 34 | 16 | 16 | 11/5/00 | 446 | 650 | 68.60% |
| 4 | Matthew Stafford | DET | 25 | 16 | 16 | 7/9/00 | 371 | 634 | 58.50% |
| 5 | Tom Brady\* | NWE | 36 | 16 | 16 | 12/4/00 | 380 | 628 | 60.50% |
| 6 | Joe Flacco | BAL | 28 | 16 | 16 | 8/8/00 | 362 | 614 | 59.00% |
| 7 | Ryan Tannehill | MIA | 25 | 16 | 16 | 8/8/00 | 355 | 588 | 60.40% |
| 8 | Andy Dalton | CIN | 26 | 16 | 16 | 11/5/00 | 363 | 586 | 61.90% |
| 9 | Ben Roethlisberger | PIT | 31 | 16 | 16 | 8/8/00 | 375 | 584 | 64.20% |
| 10 | Carson Palmer | ARI | 34 | 16 | 16 | 10/6/00 | 362 | 572 | 63.30% |
| 11 | Andrew Luck\* | IND | 24 | 16 | 16 | 11/5/00 | 343 | 570 | 60.20% |
| 12 | Eli Manning | NYG | 32 | 16 | 16 | 7/9/00 | 317 | 551 | 57.50% |
| 13 | Philip Rivers\* | SDG | 32 | 16 | 16 | 9/7/00 | 378 | 544 | 69.50% |
| 14 | Tony Romo | DAL | 33 | 15 | 15 | 8/7/00 | 342 | 535 | 63.90% |

There is an abundance of NFL Data available. We found play-by-play data for every game of the past year. Very intricate details are recorded like the wind speed and direction are recorded for every snap played over the course of the season. After looking at the play-by-play data we determined that for it to useful we would have to spend hours aggregating and analyzing the data just to get the metrics that fantasy football cares about. Once we had decided not to go with the play-by-play set we started looking for statistics that were already aggregated. We found a website that would allow us to download a file that contained every players yearly stats. Again that posed the problem of aggregation to get the data into useable form. Finally we found [www.pro-football-reference.com/](http://www.pro-football-reference.com/), which contained all the information we could hope for. We settled on year over year data for individual categories. Here is a small sample from the quarterback dataset for 2013:

# Descriptions and Relevance:

This problem relates to many topics in the course. The first thing that solving this problem does is demonstrate what students can do after taking the course. With an understanding of different clustering algorithms, classifiers, and attribute selection techniques students can find their own datasets and create models that are useful to them. One point that was stressed in the classroom was just how much time data cleaning takes up. We spent 3 hours sifting through gigabytes of data trying to find the most useful dataset. Once we had determined the most useful dataset there were many problems with the data. There were missing entries, entries that were supposed to be integers that had string values and extraneous rows. We spent another 2 hours cleaning all the data we wanted to use. An example of the necessary cleaning can be seen looking at the sample from the Data Collection section above. The QBrec column, which contains the overall record for the Quarterbacks team, comes in two different forms, and if the QB played for multiple teams the data is missing entirely. From that point we had to determine what target variable we wanted to predict. When we realized that the dataset we had collected didn’t have the appropriate target variable we had to go and find that variable from other sources and then manually insert it into our data files. The one downside of selecting fantasy football teams is that this class most likely won’t use it in the future. It is an extremely interesting problem space with many nuances and deviations that have to be considered, but you have to have an immense amount of domain knowledge and a passion for the game in order to appreciate it. Outside of this class I think an implementation of our solution could be valuable. Many people play fantasy football and it would stand to reason that a company like Yahoo! Or ESPN would pay dearly for algorithms that consistently produce better results.

# Solution:

To reiterate the problem we are trying to solve, given the input of a round number and player name we will output a yes/no answer on whether it is appropriate to draft them. To accomplish this we are going to use two different algorithms, 3-nearest neighbors and 10-means clustering. We want to use 3-nearest neighbors to classify quarterbacks because there are very few quarterbacks, which makes us think using a clustering algorithm would almost certainly over-fit the model. For the other positions we are going to use 10-means clustering. This is because Running Backs, Wide Receivers and Tight Ends score in almost identical manners and there is a large number of data points to consider. We use 10-means clustering because the average number of rounds in a fantasy draft is 10, so the idea is we create 10 clusters, order them by historical data and then check to see what cluster a particular player is in. From here the most important decision is what attributes to actually train on. We only had the chance to do meaningful analysis on Quarterback data. There were a number of features in the quarterback data set:

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Data Type** | **Description** |
| Player | String | the name of the player |
| Tm | String | three character abbreviation of the players team |
| Age | Integer | the age of the player |
| G | Integer | the number of games the player played in over the year |
| GS | Integer | the number of games the player started over the year |
| QBrec | String | the overall team record of the player for the year |
| Cmp | Integer | the number of complete passes the player threw |
| Att | Integer | the number of pass attempts the player had |
| Cmp% | String | completion percentage (Cmp/Att) |
| Yds | Integer | the total number of yards the quarterback threw for (for each completion, add up the yards gained) |
| TD | Integer | the number of touchdown passes the quarterback threw |
| TD% | String | the percentage of quarterback passes that were touchdowns (TD/Att) |
| Int | Integer | the number of interceptions thrown by the quarterback (the ball was caught by a defender, which results in a turnover) |
| Int% | String | the percentage of quarterback passes that were interceptions (Int/Att) |
| Lng | Float | the longest pass the quarterback completed |
| Y/A | Float | yards per attempt (Yds/Att) |
| AY/A | Float | adjusted yards per attempt. This is an advanced metric that gives more weight to yards gained on touchdown passes |
| Y/C | Float | yards per completion |
| Y/G | Float | yards per game played |
| QBR | Float | Quarterback Rating. Advanced statistic that takes many things into account. On a scale from 0 to 158.3 |
| Sk | Integer | the number of times the quarterback was sacked |
| SkYds | Integer | the number of yards lost from sacks |
| NY/A | Float | net yards per pass attempt. A stat that takes into account how often a QB is getting sack and how it affects total yards. |
| ANY/A | Float | adjusted net yards per pass attempt. |
| Sk% | String | the percentage of plays a QB is sacked |
| 4QC | Integer | fourth quarter come backs, the number of games that the QB led a scoring drive while losing by one score in the fourth quarter |
| GWD | Integer | game winning drives, the number of games that the QB led a scoring drive in the final two minutes of the game |

We have a number of hypotheses about the importance of different attributes. Our main hypothesis is that QBR will be the best indicator of average draft pick for quarterbacks. This is simply because it is a proprietary statistic developed by ESPN’s team of data scientists. This does pose a problem for considering historical data, since the QBR stat has only been in existence since 2006. It would stand to reason that if Tom Brady, age 37, were having a year similar to twilight years of John Elway’s Career then you would probably want to draft him early (In 1997-98 John Elway won back-to-back Superbowls at the age of 38 and 39, with dominating performances throughout the year). On the other hand if Tom Brady were having a year similar to the twilight years of Brett Farve’s career (who struggled mightily) then you would most likely not want to draft him. Regardless of the historical implications, QBR is one of the most widely used statistics to compare current NFL Quarterbacks, so it would make sense to assume it would be very predictive. Another hypothesis we had is that “Fourth Quarter Comebacks” and “Game Winning Drives” would be very predictive as well. It is very hard to determine the “clutch” factor of a Quarterback and we were thinking that maybe these stats could help (clutch factor is widely used to refer to how well a player performs under immense pressure; For example Jim Kelly of the Buffalo Bills was a very clutch player, while Tony Romo of the Dallas Cowboys is very un-clutch). The final hypothesis we had was that sacks, sack yardage and sack percentage would be important attributes considering the nature of sacks. If a Quarterback gets sacked it usually ends up in what is called a “long yardage” situation, meaning you have longer than 10 yards to go to get a first down. Long yardage situations are critical because it causes the Quarterback to throw passes that maybe he would think twice about under normal circumstances. Overall we thought sacks would be a very predictive, non-obvious attribute.

# Results:

The results of our experiment were mixed at best. We ran Principle Components Analysis on the Quarterback data to try and determine the most important attributes. The following is the percentage of variation of the principle component vector caused by each attribute (if you sum them up they add to one):

|  |  |
| --- | --- |
| **Attribute** | **Variation in PC Vector** |
| QBR | 0.421 |
| TD | 0.238 |
| Int | 0.082 |
| Att | 0.048 |
| Cmp | 0.045 |
| Cmp% | 0.037 |
| Yds | 0.026 |
| TD% | 0.023 |
| G | 0.018 |
| GS | 0.018 |
| Age | 0.012 |
| QBrec | 0.007 |
| Y/G | 0.007 |
| Y/C | 0.006 |
| AY/A | 0.004 |
| GWD | 0.003 |
| 4QC | 0.002 |
| Int% | 0.001 |
| Lng | 0.001 |
| Y/A | 0.001 |
| Sk | 0 |
| SkYds | 0 |
| NY/A | 0 |
| ANY/A | 0 |
| Sk% | 0 |

As we expected the QBR is the most important attribute by far. This makes perfect sense given how popular a stat this is (data-miners would have realized by now if it were a poor indicator of success). From there most of the results make sense. You score points by throwing touchdowns so it would make sense that it is the second most important attribute. The next few attributes are all scoring methods that received weights that intuitively make sense to us. We were disappointed to see that “Fourth Quarter Comebacks” and “Game Winning Drives” were not strongly represented in the principle component vector. However they do carry some importance and these results make sense (if it was as easy as that to determine how “clutch” a player is then there wouldn’t be constant debates about it in the media). The most disappointing result was the importance (or lack thereof) of sacks. We really thought that sacks contained “hidden yards” that most models wouldn’t account for. It turns out that most models don’t account for them because they are relatively unimportant. While we still think sacks carry hidden information, it is probable that the information is a better predictor of team success as a whole rather than the individual success of the Quarterback. It is probable because sacks aren’t caused by the Quarterback, they are caused by myriad of different things including the offensive play call, the defensive play call, the talent level of the offensive and defensive lines, etc….

# Conclusion:

Overall there is plenty of data to suggest that player statistics can be used to predict a fantasy draft position. However reflecting on the original implementation plan we would say much more work is needed. We should build models for many different aspects of the player and then combine the results of those models into a decision. We could build models to give players a “clutch factor”, to predict the likelihood of injury throughout the season, to predict the likelihood of team success (the better your team the more likely you are to score points), etc…

We learned many things about data mining during this project. One thing that stood out is how easy it is to do PCA analysis. During class it was stressed that despite the underlying beauty of the math required to do PCA, it really is just a function call for most programmers. This was the case for us. Another thing we learned is how tricky data cleaning can be. Despite our best efforts there were errors in the data that made analysis difficult. Overall this was a great project for us as a team and given our interest in data mining and fantasy football, we will probably continue to develop our project plan.