Optimizing Preseason Fantasy Football Rankings

John O'Hollaren (jpo4)
Department of Electrical and Computer Engineering

Abstract

Fantasy football is a fast growing, multi-billion dollar industry centered around projecting which players will perform the best over the course of a season. Points are awarded when a player accomplishes certain tasks, such as scoring a touchdown, rushing for yardage, or catching a pass. Teams are picked at the beginning of the year, typically through a snake style draft where players are taken one after another. Correctly predicting which players will play well greatly improves your team's year-long performance, whereas wasting a high pick on a low achiever can ruin your season. This paper focuses on an optimal strategy for predicting which NFL wide receivers will have the best fantasy football season using data available prior to the beginning of the season.

1 Introduction & Related Work

1.1 Data Summarization

My data includes NFL statistics every year from 2007 to the present. For each year, these statistics include: receiving yards, receiving targets, receiving catches, receiving touchdowns, rushing yards, rushing touches, rushing touchdowns, fumbles, number of games played, and team. I manually compiled this data from several sources, and I have it for about 1000 players for the years 2007 to present.

I also collected preseason ranking data from the experts at ESPN and Yahoo for 2007 to the present. This is my target - these are the rankings that I want to improve upon. These are also made available prior to the beginning of the season, so I can also use them as a feature input. This data is not readily available, to get this data, I parsed HTML from old versions of the ESPN and Yahoo sites.

1.2 Conditions

For this problem, I will not consider all non-rookies. Adding rookies would involve delving into data from college and from the NFL combine, exponentially increasing the size of the problem. Furthermore, rookie performances are extremely dependent on factors such as team depth and starter injuries, factors that are difficult to include in a mathematical model.

2 Models and Methods

2.1 Regression

A large amount of data is missing in this dataset. For example, predicting a 2nd year player's performance is just as important as predicting a 10 year veteran's performance, but for the veteran I have 10 years of data to rely on whereas for the younger player I do not. Therefore, I cannot use the same features for every player. However, if I use a measure of 'slope' of change from one year to the next for various stats, I can set all missing years to 0. This will set a linearly increasing prior.

2.2 Features

My features for predicting fantasy points (FP) in year Y are:

FP_{Y-1}
$receiving\ yards_{Y-1}$
$receiving\ catches_{Y-1}$
$receiving\ tds_Y - 1$
$rushing\ yds_{Y-1}$
$rushing\ touches_{Y-1}$
$rushing\ tds_{Y-1}$
$fumbles_{Y-1}$

Then I also want to include how these changed over time:

$\Delta 1_{receiving\ yards} = receiving\ yards_{Y-1} - receiving\ yards_{Y-2}$
$\Delta 1_{receiving\ tds} = receiving\ tdsY - 1 - receiving\ tdsY - 2$
$\vdots$
$\Delta 2_{receiving\ yards} = receiving\ yards_{Y-1} - receiving\ yards_{Y-3}$

Then per the above, if a player only played 2 years, then $\Delta 2_{receiving\ yards} = receiving\ yards_{Y-1} - 0$. I currently run a linear regression on this model, where X is a non-hidden (observed) matrix of all my data per player, and β are my weights from the training data (previous years). $Y_{predict}$ is my output of predictions for the next year, and it is equal to $X_{test} * \beta$. $Y_{predict}$ can be interpreted as a confidence in how well a player will play - the higher it is, the higher that player should be drafted in the fantasy football draft.

2.3 Mixed Model

I am also going to look at breaking my data into different groups. There are different types of receivers (wide outs, slot receivers, etc), so I may be able to improve the accuracy of my predictions by modeling the groups separately. I will look at three methods for this: manually setting my groupings, K-means on some statistics, and also semi-supervised clustering. The end result will be a different β_k for each mixture, where I may have latent hidden groupings, or I may not, depending on which method I use.

3 Results

4 Conclusions

5 Bibliography