CS 221 Final Project Proposal

cmayer

October 2014

1 Problem Description

Fantasy Football (FF) is one of the most popular hobbies in the country. FF participants (called owners) make weekly decisions about which football players to put in their starting lineups with the goal of maximizing their team's overall performance, measured in FF points. Even though massive amounts of detailed data is available for every player, owners usually make decisions based on a few indicators and gut instinct. For our project we plan on building a Fantasy Football lineup predictor that, given a set of features about the past performance of a football player, will try to predict his performance in a given game.

2 Prior Work

Most of the prior work on this problem has been informal – FF participants privately employing machine learning methods to better the performance of their own teams. Of those such attempts with available documentation, most have used unsupervised methods and ensemble methods on expert projections.

3 Data

Pro-football reference provides CSV formatted statistics from every major category for every game in the last 50 years. We would limit our training and evaluation to the top 100-200 players, since players with limited playing time are unlikely to be on a fantasy roster and have limited signal to learn from. Our input data would be a feature vector based of various player statistics:

Player	Yards Last Game	Opponent	Quarterback Rating	Age
Randy Moss	89	NYJ	89.2	31

FF uses a scoring function to combine different statistical categories into a weighted total score. Prior to training we would choose a commonly used scoring function to produce our expected output.

$$FFScore = Yds/10 + Touchdowns * 6 - Fumbles/2$$

Our regression model would then directly predict the FF score for a player in a given week.

4 Baseline and Oracle

Historical averages provide a simple and intuitive baseline model for predicting a player's performance in a given week. Given a player's statistics over a number of games played, the historical average baseline model simply outputs the average of the player's performance as a prediction for future games. This approach represents the The closest thing to an oracle is the service provided by major FF leagues, like Yahoo, ESPN and CBS, which similarly projects player performance on a game by game basis. The system used to develop these projections is a black box, but appears to be a combination of machine learned features and expert opinions. Our goal will be to try and produce similar performance using only machine learning.

5 Approach

We will explore the use of different regression models, starting with linear regression and possibly moving to neural networks if we need more expressive power. The majority of the work and performance gains is expected to come from feature engineering: finding which pieces of data, and combinations of pieces of data, provide the best signal for predicting player performance.

6 Evaluation

With such a large corpus of available data, evaluation should be straightforward. We can train on a collection of historical data and previous games from the current year, then predict the performance for a player each week and use the actual statistics after the game to evaluate our predictor.