Using Supervised Machine Learning Algorithms to Predict Fantasy Football Player Scores

Garrett Johnston, Alex Wood
University of California, Los Angeles
(gjohnston@ucla.edu, alex.wood@cs.ucla.edu)

December 11, 2015

Abstract

The appeal of NFL fantasy football is its stochastic nature, which has launched a gambling industry around it and inspired experts to attempt to analyze games and predict future score outcomes. The aim of this project is to acheive those goals by utilizing supervised machine learning models, specifically linear regression and support vector machines. With these models, we also explore the importance of a number of features selected based on domain-specific knowledge. We evaluate the accuracy of our models through cross validation as well as comparison to expert predictions from ESPN. Although the SVM models proved to be somewhat ineffective, they outperformed the ESPN projections for all three player positions modelled, Quarterback, Running Back, and Wide Receiver.

1 Motivation

NFL Fantasy Football has become a national sensation in the community of sports gambling and betting. With one-week fantasy sites like FanDuel and DraftKings cashing out millions of dollars in winnings per week, it is economically valuable to predict the outcome of individual players' statistics in a game. It is also an interesting challenge to create mathematical models of human strategy and physical ability, as well as attempt to overcome the seemingly inherent randomness of the sport.

2 Background

There are expert analysts and machine learning techniques that have been developed that attempt to predict the outcome of games [1]. There are also attempts to predict fantasy scores using machine learning [2]. While these approaches have proven to be successful, our proposed project will attempt to discover a number of different features than those used in the previous work. We include fine grain teammate statistics, as well as the opponent team's average defensive statistics.

3 Data

All of the football statistics that we needed could be found for free on http://www.pro-football-reference.com. We wrote a scraper to collect all game data from 2011-2014 for every active franchise during those years. Data included statistics about individual players, and statistics for the team as a whole. After storing the data into a MongoDB NoSQL database, we used it to generate feature vectors based on our domain-specific knowledge. We grouped our data into collections of players, containing relevant information on their individual statistics in the games they have played, and teams, containing data pertaining to team performance.

Each label represents a player's fantasy score for a single game, defined by a simple and standardized function of the player's individual statistics that game (See Equation 1). Typical fantasy scores are in the range of 0 - 35. As the most important players (i.e. those who contribute the most fantasy points) are the quarterback, running back, and wide receiver, we have chosen to train models only for each of these positions. The feature vector contains pre-game knowledge about the player, the team, and the opposing team, as follows:

- **Individual player statistics:** The obvious predictor of a player's fantasy score is their performance in prior games. We include the player's average statistics over the past six games, as well as over the entire season up to this game. The set of individual statistics chosen are the relevant measures of performance for that player's position. For instance, the features for a quarterback will consist of passing yards and passing touchdowns, while a wide receiver's features will be based on their receptions and rceiving touchdowns. It is worth noting that all of these offensive skill positions are capable of attaining rushing yards and rushing touchdowns, thus we include these in the features for all players. As the set of features will be different for each position, we need to train separate models for each.

- **Team statistics**: Football is the ultimate team sport. A single player's performance is highly dependent on the performance of his team as a whole. With this, the next part of the feature vector is composed of team performance measures. Such measures include total yards gained per game and points scored per game. Similar to the individual player features, we took the average of the team statistics over the six most recent games, as well as over the entire season up to this game. The rationale behind choosing the six most recent games is that a team's recent performance is often indicative of their future performance.

- **Opposing team's statistics:** A player can only play as well as the defense will allow him. Therefore, including the abilities of the opposing team is imperative to correctly predicting fantasy output. With this, we have computed similar measurements to team statistics, but with respect to the defensive statistics of the opposing team. Some features are the defense's yards allowed, touchdowns allowed, and points allowed averaged, again, over the span of the six most recent games as well as the over the entire season preceding this game.

Since we rely on past data to construct the feature vector for a given player and game, we are unable to include data for rookies and new players that have no past NFL experience. We address an option to work around this in the future works section.

$$score = pass\,yards/25 + 4 * pass\,touchdowns - 2 * pass\,interceptions + rush\,yards/10$$
$$+6 * rush\,touchdowns - 2 * fumbles\,lost + rec\,yards/10 + 6 * rec\,touchdowns$$

Equation 1: Function to calculate a player's fantasy score

4 Methods

4.1 Linear Regression

Our first approach was to use the simplest regression model for predicting continuous labels. We initially evaluated this model by partitioning the training data into sub-training and sub-test sets, attempting to predict the sub-test sets as if they were unseen data. The evaluative measure was the absolute value of the difference between the predicted values and the actual values of the label. After this, we used 5-fold cross validation to ensure our predictive procedure was sufficient.

4.2 Support Vector Machines

Our second approach was a regessor model using SVM with an RBF kernel. We used cross-validation to discover the most predictive model for different values of C, the penalty term of the error, and Γ, the kernel coefficient for RBF. We also used SVM as the model provided to a feature selection algorithm, Recursive Feature Elimination with Cross Validation (RFECV).

5 Evaluation

5.1 Linear Regression

The following are the average absolute prediction errors for models predicting fantasy scores for Running Backs, Quarterbacks, and Wide Receivers respectively, using linear regression and 5-fold cross validation:

- Quarterbacks: 5.95
- Running Backs: 4.90
- Wide Receivers: 4.28

5.2 Support Vector Machines

We measured the average absolute prediction errors using various values of C for SVM with an RBF kernel, but found that the impact of C was negligible. Therefore, C was held at its default value for our SVM library. For various values of Γ, the following are the average absolute prediction errors for our SVM models corresponding to different player positions:

We used SVM with out best observed value of $\Gamma = 4^{-7}$ to represent the performance of the models against our test data over the range of fantasy scores. Plots of the absolute value of the error for each position are shown in Figure 1.

Γ	4^{-10}	4^{-9}	4^{-8}	4^{-7}	4^{-6}	4^{-5}	4^{-4}	4^{-3}	4^{-2}	4^{-1}
Quarterback	6.16	6.13	6.12	**6.09**	6.09	6.20	6.28	6.29	6.29	6.29
Running Back	5.16	4.94	**4.87**	4.92	5.03	5.26	5.42	5.43	5.43	5.43
Wide Receiver	4.25	4.15	**4.13**	4.14	4.20	4.34	4.51	4.55	4.55	4.55

Table 1: SVM absolute errors for different values of Γ

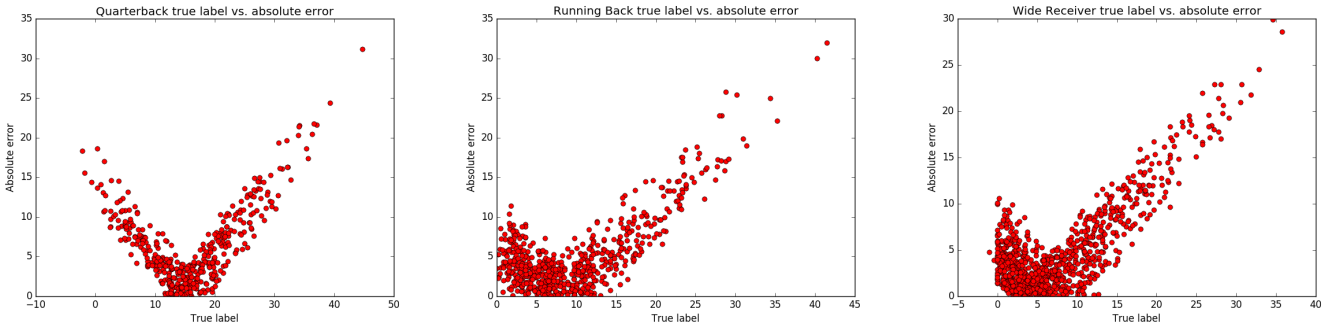

Figure 1: SVM with RBF kernel, with $\Gamma = 4^{-7}$

5.3 Feature Selection

An interesting result of this project is discovering the features that are most impactful on the resulting prediction accuracy. Using an SVM model with a linear kernel, we applied RFECV with 5-fold cross validation, and found the ranking of features for the Running Back model shown in Table 2.

Using the recommended features from the algorithm, the new average absolute prediction errors using an SVM model with an RBF kernel having $\Gamma = 4^{-7}$ for the model of Running Backs is 4.70, an improvement on the previous best of 4.87. The recommended features provide a slight improvement over the original features.

5.4 Comparison with Expert Projections

One of the major initiatives of this project is to provide a fantasy score prediction algorithm that performs better than some or most of the currently available expert predictions. Some well respected sources in the fantasy football community are NFL.com fantasy, Yahoo fantasy, CBS fantasy, FFtoday.com, and ESPN fantasy. We chose to compare our model's predictive power with noted sports powerhouse, ESPN. As shown in Table 3, our model performs better in prediction error for all three positions.

	Quarterback	Running Back	Wide Receiver
ESPN prediction error	6.47	5.25	5.38
Our SVM prediction error	6.09	4.70	4.13

Table 3: ESPN prediction error vs. our SVM prediction error

Feature name	Feature importance
starting	1
rush yards (past 6 games)	2
rush attempts (past 6 games)	1
rush touchdowns (past 6 games)	1
fumbles (past 6 games)	1
fumbles lost (past 6 games)	1
rush yards (past 16 games)	1
rush attempts (past 16 games)	7
rush touchdowns (past 16 games)	1
fumbles (past 16 games)	1
fumbles lost (past 16 games)	1
receiving yards (past 6 games)	1
receptions (past 6 games)	8
receiving targets (past 6 games)	1
receiving touchdowns (past 6 games)	1
team total yards (past 6 games)	1
team passing yards (past 6 games)	3
team rushing yards (past 6 games)	4
team points (past 6 games)	5
team turnovers (past 6 games)	6
opponent points (past 6 games)	1
opponent total yards (past 6 games)	1
opponentpassing yards (past 6 games)	1
opponentrushing yards (past 6 games)	1
opponent forced turnovers (past 6 games)	1

Table 2: Ranking of feature importance (1 is most important) using RFECV with 5-fold cross-validation and SVM linear kernel

6 Software Tools

Our primary language for extracting and generating data is Python. The supervised learning models used in our methods are from the scikit-learn Python library. Our data scraped from the web is stored using MongoDB.

7 Conclusions & Future Work

Fantasy football prediction is a new and exciting application of machine learning techniques. The current state of the art is far from perfect, but there it is clear from the results that there is an underlying correlation between past statistics and future statistics.

Given the possible range of fantasy scores, the accuracy of our models is far from ideal. However, our sample experiments have shown that our fantasy estimates can compete with the experts

in fantasy football prediction. Our results also show that certain positions tend to be easier to model. In this case, Quarterback performance was harder to predict. Additionally, our results from experimenting with tuning the SVM algorithm show that changing the value of C had little to no effect on the overall accuracy of our model. This is likely due to the limited set of features we selected.

From the results of the feature selection algorithm, we found that individual statistics and opponent team's statistics were the most relevant features in the outcome of the player's fantasy score. The players own team statistics were surprisingly selected as less important than the other features. Additionally, it is unclear whether or not the most recent games are more influential than cumulative statistics over the entire season. Future work would involve comparing the importance of statistics over different periods of time.

One of the limitations of our data, was restricting any new players or college players, as they have no NFL data. Future work could include extrapolating past statistics from their college game history, normalizing to NFL standards. Other future work could involve more complex features, such as a similarity metric that identifies fantasy performance in games with the most similar conditions (e.g. weather, time of day, home vs. away, week in season, teammates playing, etc.).

References

[1] Jim Warner, *Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line*, `http://www.cs.cornell.edu/courses/cs6780/2010fa/projects/warner_cs6780.pdf`

[2] Roman Lutz, *Fantasy Football Prediction*, `arXiv:1505.06918 [cs.LG]`