Predicting Fantasy Football - Truth in Data
Matt Bookman
December 14, 2012

1 Introduction

The study of sabermetrics is defined by Bill James as "the search for objective knowledge about baseball."[1] Via data analysis, sabermetrics seeks to dispel or validate traditional measures used to construct a winning baseball team. The goal of this paper is to begin the search for objective knowledge about fantasy football.

Yahoo Sports provides an online fantasy football league to its user base. End users create leagues and then those individuals construct a roster of real NFL players each week for head-to-head competition. NFL players are assigned “fantasy points” based on their achievements in each football game, and each head-to-head matchup produces a winner and loser based on the respective teams’ accumulated fantasy points. Fantasy points are assigned based on elements such as rushing yards, receiving yards, passing yards, and points scored.

Yahoo Sports also provides "Projected Points" from a 3rd party[2] for each NFL player prior to a game, to help end users better assemble a team lineup for each week. Along with the Projected Points come branded "News and Notes" and "Scouting Reports" to help one make informed decisions in filling out a weekly lineup.

Personal experience has indicated dubious value for the Projected Points and Scouting Reports. The objective of this project is to produce more useful predictions of fantasy points and to glean the value of advice such as a team playing at home or on the road or a quarterback facing a strong passing defense.

Results of linear regression and ranking SVM machine learning predictions are compared against the Yahoo Sports Projected Points as well as a baseline random predictor.

2 Data

To generate predictions, player and team data was sourced by scraping web pages from ESPN.com’s weekly NFL statistics and parsing the HTML. Player statistics[3] and game results[4] (team statistics) were source for Week 1 through Week 10 of the 2012-2013 NFL Season.

Earned Fantasy Points and Projected Fantasy Points were sourced by scraping web pages from the Yahoo Fantasy Football web site.[5] Fantasy data was sourced for quarterbacks, running backs, wide receivers, and tight ends for Week 5 through Week 11 of the 2012-2013 NFL season. Fantasy Projected Points were sourced for the top 50 for each position. Fantasy Points (results) were sourced separately for the top 25 at each position.

For this project, focus was narrowed to predictions on the quarterback position. The quarterback was chosen for a number of reasons, chiefly:

1. Quarterbacks provide the most consistent dataset. There is typically one quarterback who plays the entire game, and so the projected versus actual data is most complete. 25 training elements were produced for each week.
2. Quarterbacks acquire both passing and rushing fantasy points. Not all quarterbacks generate rushing yards, but enough do to force the inclusion of rushing statistics in the analysis of quarterbacks. This adds to the interest of the problem.

The initial feature set used for machine learning predictions can be seen in Table 1. All data statistics are treated as per-game averages.

Individual Player	Opposition Defense	Game Specifics
Passing Completions	Passing Yards	Home/Away
Passing Attempts	Yards Per Pass	
Passing Yards		
Passing Touchdowns	Rushing Yards	
	Yards Per Rush	
Rushing Attempts		
Rushing Yards	Points	
Yards Per Rush	Total Yards	
Rushing Touchdowns	Yards Per Play	
Sacks	Sacks	
Interceptions	Interceptions	
Fumbles	Fumbles Recovered	

Table 1: Machine Learning Feature Data

3 Preliminary Data Analysis

Figure 1 shows a histogram of all quarterbacks' per-game fantasy points and a histogram of the Yahoo Sports prediction errors.

While the mass of the errors in Projected Points is in the -7 to 7 range (standard deviation: 7.0502), the average Fantasy Points is just over 17, making such an error range quite large. In fact a naive predictor that blindly predicts the mean produces a slightly better average absolute error (5.492 vs 5.5374) and a slightly narrower standard deviation (7.0464).

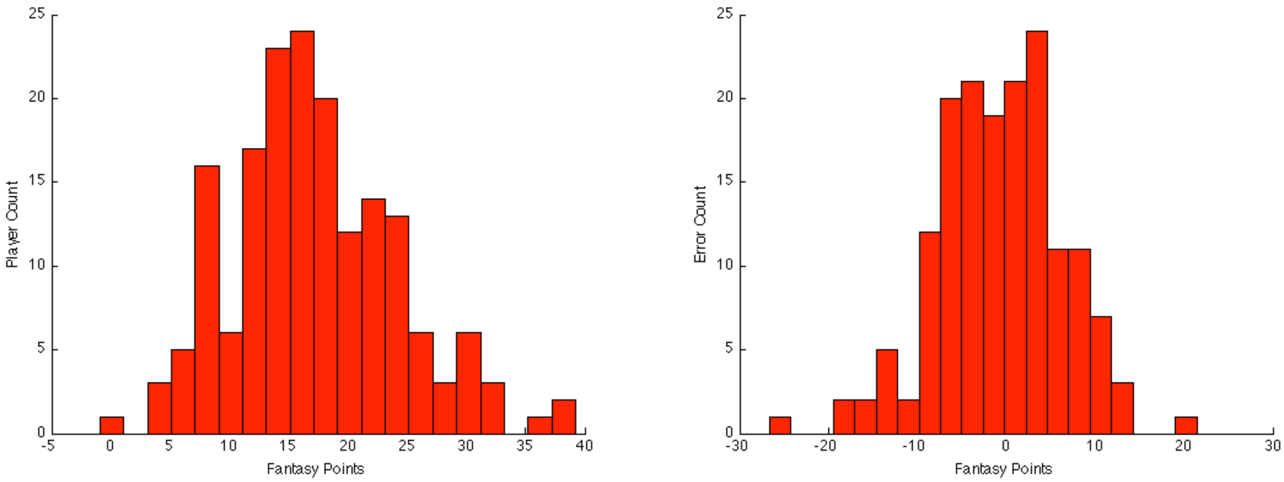

Figure1: (**a**) Weekly Fantasy Points for NFL quarterbacks (**b**) Prediction Error in Projected Points

However, the value of a prediction for building a team is not necessarily in getting close to the actual points. When filling out a roster in a given week, more valuable would be a **ranked** list of available players for a position. The value is in the ordering of projections, not the accuracy of the values.

Thus the metric used for comparison of the machine learning projections to the Yahoo Projected Points is a Pearson Rank Correlation coefficient.

4 Data Preparation

An attempt was made to capture two commonly held approaches to predicting how a player will perform: 1) How has the player performed over the course of the season, and 2) How has the player performed recently. As feature input for each training element, the individual player weekly values were included as:

1. Previous game's totals
2. Last 4 games' average
3. Season average

Opposition defense statistics were included as season averages only.

A training element thus consists of 44 feature fields (3 * 11 individual + 10 defense + 1 home/away). The target value for each record is the Fantasy Points the NFL player accrued for the given week.

5 Machine Learning Methods

5a. Linear Regression

Linear regression over training elements for quarterbacks was performed using the Matlab `regress` function. Predictions were then made by computing `X * w`. `X` is a matrix such that each row is a training element and the columns are the input features (plus constant). `w` is the learned coefficient vector (including constant term).

To be of true value and not just academic interest, one must be able to make predictions for a coming week based only on past statistics. Thus the procedure followed is:

```
For each WEEK: MIN_WEEK + 1 through MAX_WEEK
     Train on data for MIN_WEEK through WEEK - 1

     Test on data for WEEK producing projected Fantasy Points

     Convert linear regression WEEK projected points into a vector of ranks
     Compute Pearson's rho for linear regression rank versus actual rank

     Convert Yahoo WEEK projected points into a vector of ranks
     Compute Pearson's rho for Yahoo rank versus actual rank

     Compute Pearson's rho for random permutation (1 to TEST_SIZE) versus actual rank
```

5b. Ranking SVM

Support Vector Machines (SVMs) are one of the most powerful tools available in machine learning today with the ability to trade off training error predictions for better generality (larger margins), and the ability to use kernels to explore higher dimension feature vectors. Efforts to produce better web page results has led to a technique of applying SVMs to make ranking predictions[6]. For each pair of training elements the rank ordering between them becomes a constraint of the objective function.

The SVM^{rank} software was used to make ranking predictions in a similar fashion to linear regression. Six data files were produced representing ranked training data for weeks: 5, 5 through 6, 5 through 7, ... 5 through 10. Six data files were produced representing test data for weeks 6, 7, ... 11.

One significant difference between the SVM results and linear regression is that the SVM results are of value strictly as ranking and not as predictors of specific projected Fantasy Football points.

SVM^{rank} was run with all default parameters, notably the SVM parameter "C". The C value allows one to make trade-offs between having training elements be in error with increasing the size of the separation margin. Raising this value from its default of 0.1 only produced inferior results.

Attempts were made to run SVM^{rank} with non-linear kernels, but the performance was so poor as to make those options unusable.

6 Preliminary Results

Early results of linear regression when limited data was available produced poor results. An approach was taken to make random 70%/30% splits of the data into training and test sets in an attempt to extract a best model. This approach did produce better results when limited data was available.

However this approach turned out to simply be an ad-hoc way of finding the most predictive features. The results of `regress` provide confidence bounds for the generated coefficients. These results were used to eliminate the least informative features. It was ultimately observed that specific weeks (training through week 6 and training through week 8) had models that predicted better than others over all subsequent weeks. Both models learned had eliminated all but 4 of the original 44 fields:

Week 6	**Week 8**
Pass Yards (last 1 game)	Pass Yards (last 1 game)
Pass Yards (last 4 games)	Pass Yards (season)
Rush Yards (last 4 games)	Rush Yards (last 4 games)
Defense Passing Yards (season)	Defense Passing Yards (season)

Empirical testing demonstrated little difference in prediction results for the two choices (mean 0.01 per week), but the better of the two (Week 8 features) were chosen as the feature set used for **all** predictions.

7 Final Results

The first prediction week (6) had effectively no correlation in rank predictions (LR: -0.0090, SVM: -0.0045) which is likely due to having a single week of training data. For four of the five subsequent prediction weeks, both linear regression and SVM predictions were superior to Yahoo Projected Points and often by a significant margin (Figure 2).

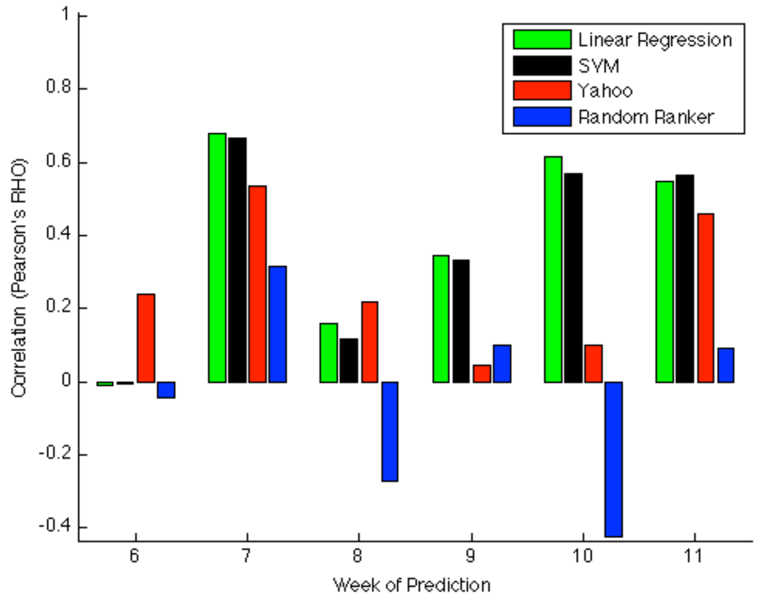

Figure 2: Rank Correlations for all prediction methods Week 6 through Week 11.

In the four weeks that linear regression and SVM ranking correlations were higher than the Projected Points, the average correlations were 0.26 (LR) and 0.25 (SVM) higher. In the single week they were lower it was by 0.06 (LR) and 0.10 (SVM) respectively.

	LR	SVM	Yahoo
Weeks 7, 9, 10, 11	0.5457	0.5322	0.2616
Week 8	0.1571	0.1176	0.2174

Table 2: Mean rank correlation over weeks when machine learning was superior/inferior to Yahoo

The results of a random ranking predictor are shown in Figure 2 to demonstrate that the machine learning and Yahoo predictions are better than random, and also that one can be fooled (see Week 7, 9, and 11) if not careful.

A critical step in achieving these results was the reduction of features discussed in the previous section. Figure 3 shows the result of better feature selection on both linear regression and SVM. When trained on all features, linear regression predictions were better than Yahoo in only 2 of the 5 key prediction weeks. SVM predictions were better than Yahoo in only 1 of the 5 key prediction weeks. With improved feature selection, mean RHO for linear regression improved from 0.2734 to 0.4680 and for SVM improved from 0.1723 to 0.4493.

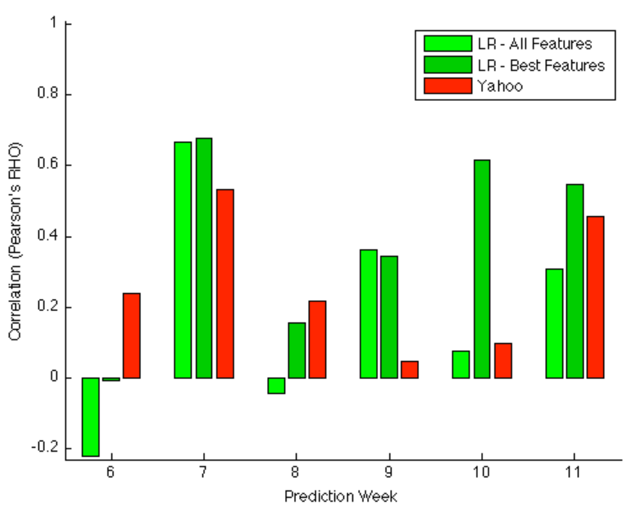

Figure 3a: Linear regression predictions with all features and with only the most predictive features.

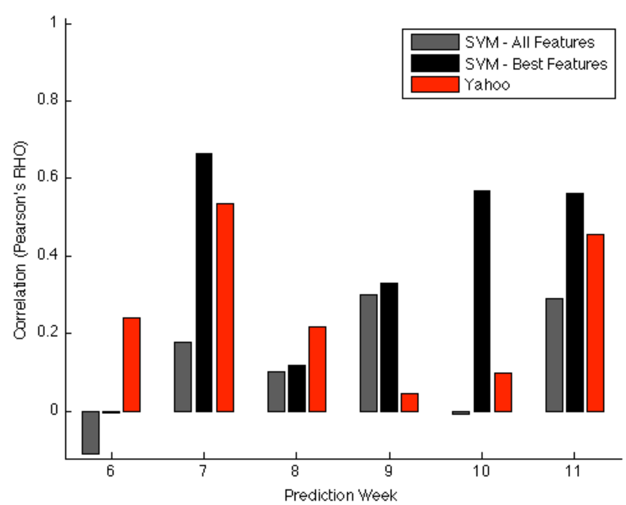

Figure 3b: SVM ranking predictions with all features and with only the most predictive features.

Table 3 shows the mean contributions of each of the 4 remaining features. An interesting result is that on average a more significant contribution to a quarterback's fantasy points (and wider variation) comes from the quality of the opposition passing defense than from the quarterback's own statistical history. Predicting rank only on the defense yields a superior correlation coefficient (0.3813) to individual passing yards (0.1582).

	Pass Yards (previous game)	Pass Yards (season)	Rush Yards (last 4 games)	Defense Passing Yards (season)
Mean/STD	-3.7731/1.4412	12.1701/2.2291	0.2729/0.4242	16.8117/2.7755

Table 3: Average contributions to linear regression predictions for each feature. Including the constant term (-8.4397 mean) yields a mean prediction of 17.0419 fantasy points.

The primary objective of this project was to produce superior rankings. However linear regression numerical predictions of the Fantasy Football points are not devoid of value. When compared to Yahoo's projections, linear regression results are again superior as shown in Figure 4.

Linear regression average prediction error was 4.97 versus Yahoo's 5.49 and is better in all weeks. However, that the linear regression mean error is less than Yahoo in Week 6 illustrates the point that on its own, minimizing mean prediction error may not achieve the desired result as rank correlation for Week 6 was virtually 0.

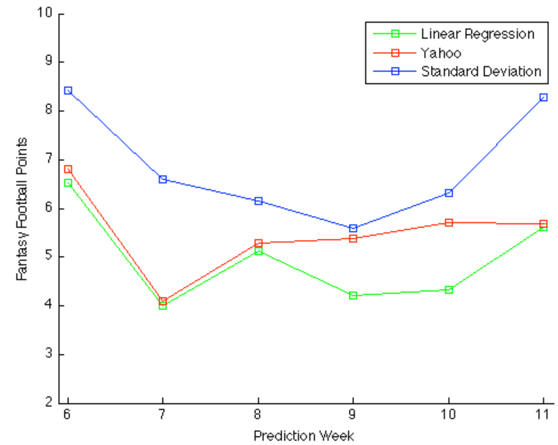

Figure 4: Linear regression has mean prediction error less than Yahoo and well within standard deviation for points

8 Conclusion

As with all sports, the games are played on the field and are not simply simulations based on statistics. One can never expect perfect predictions. However, this project has demonstrated that machine learning techniques can produce results of significant quality in predicting the rankings of NFL quarterbacks for Fantasy Football. Both machine learning approaches applied here produced consistently better rankings than Yahoo's Projected Points.

While both linear regression and ranking SVM produced strong results, on the whole, the results of SVM^{rank} were somewhat disappointing. A significant draw to using an SVM is the ability to use different kernels and explore higher dimensional mappings of the original feature space. The SVM^{rank} implementation in practice did not allow for using anything but a linear kernel which made the results almost indistinguishable from linear regression.

There is much still to be explored to build on the results achieved here. A preliminary analysis suggests building a mixture model with two different categories of quarterbacks: 1) "the running quarterback" (there are 7) and 2) the "pocket quarterback" (there are 33). These two categories may be better modeled by two different sets of features and coefficients.

In the end, the large numbers of features producing inferior results demonstrates that the right data and right model are critical and many elements perceived as relevant by "expert" scouting reports are in fact, noise.

9 References

[1] Grabiner, David J. "The Sabermetric Manifesto". The Baseball Archive. (http://seanlahman.com/baseball-archive/sabermetrics/sabermetric-manifesto/)

[2] Yahoo Help. "Understanding Projected Points in Fantasy Football" http://help.yahoo.com/kb/index?page=content&id=SLN6625

[3] ESPN NFL Weekly Leaders - 2012, http://espn.go.com/nfl/weekly/leaders/_/type/

[4] ESPN NFL Schedule - 2012, http://espn.go.com/nfl/schedule

[5] Yahoo Fantasy Sports, http://football.fantasysports.yahoo.com (note: user account required)

[6] Joachims, Thorsten. Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002.