Building an NFL Performance Metric

Seonghyun Paik

Department of Electrical Engineering, Stanford University

Introduction

In this project, I focus on a supervised learning on which statistic category of National Football League (NFL) is more related to a team's score, and finally a win. In addition, the result is compared with traditional metric. (passer rating)

Data Acquisition and Processing

I. Raw data acquisition

Individual NFL game statistics in 2013 is obtained from https://www.statcrunch.com/app/index. php?dataid= 1903988. Since each 32 teams will have 16 games (except post season), the raw data contains data for 256 games.

II. Selecting and adding features

Two features are selected for my prediction variables - the scored points for each team and win/loss of the games. I've included additional features (YDS/A, PCT) since they could provide more intuitive analysis and traditional metric is also based on them.

III. Selecting training and test samples

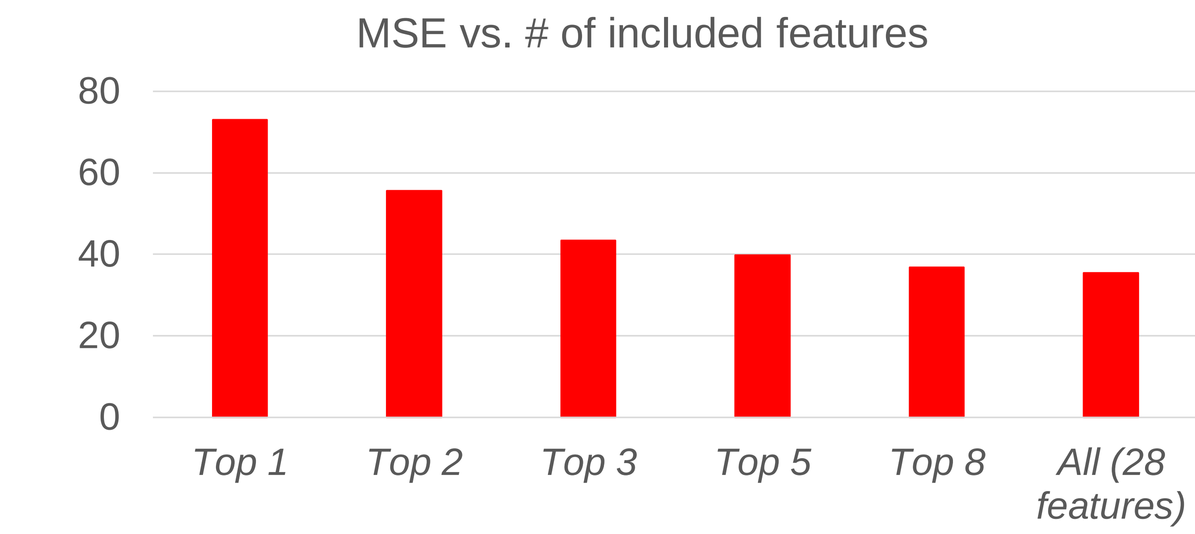
From the 512 score samples, I set aside 400 samples for training and validation, and the remainder (112 samples) for testing. For prediction of win/loss, I selected 200 game results for training and validation, and remaining 56 results are reserved for testing.

Methodology

I. Feature Selection

I ran sequential forward-based feature selection on 400 training examples. Features were selected based on their mean-squared error using 5-fold cross-validation. The top features after feature selection are below.

1	1Down	5	RushY/A
2	PassY/A	6	Yard Loss from Sack
3	TO take away	7	Opp. Penalty Y
4	TO give up	8	Penalty Y

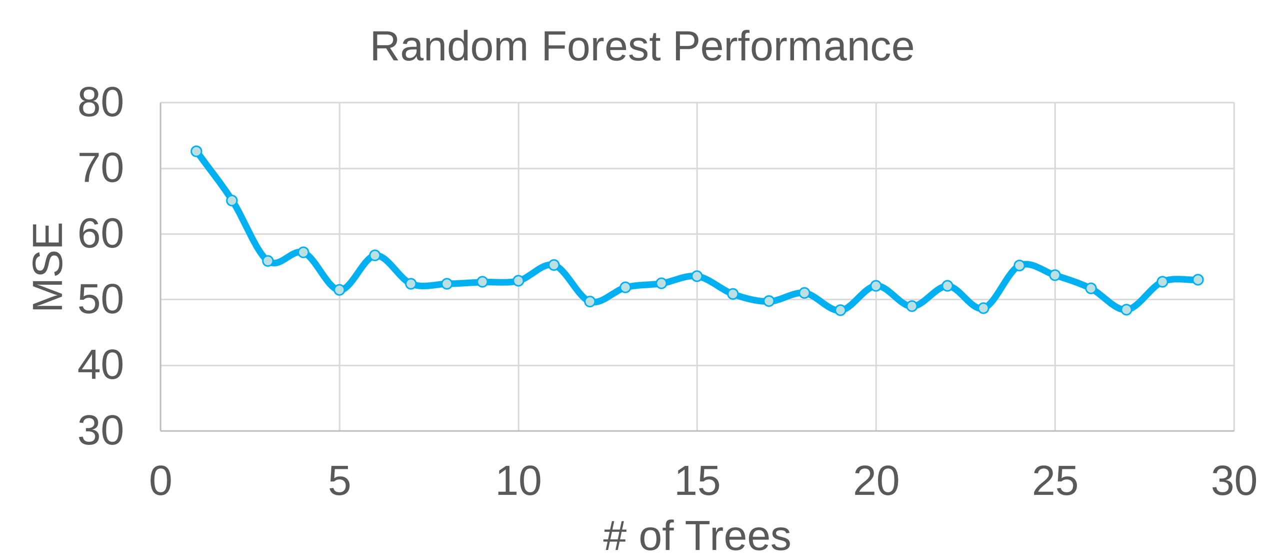


II. Linear Regression

Using obtained features, ensemble method using weighted sum of linear model (inversely proportional to training MSE) was used. Ridged regression and weighted linear regression has been also tried, but found no noticeable performance enhancement.

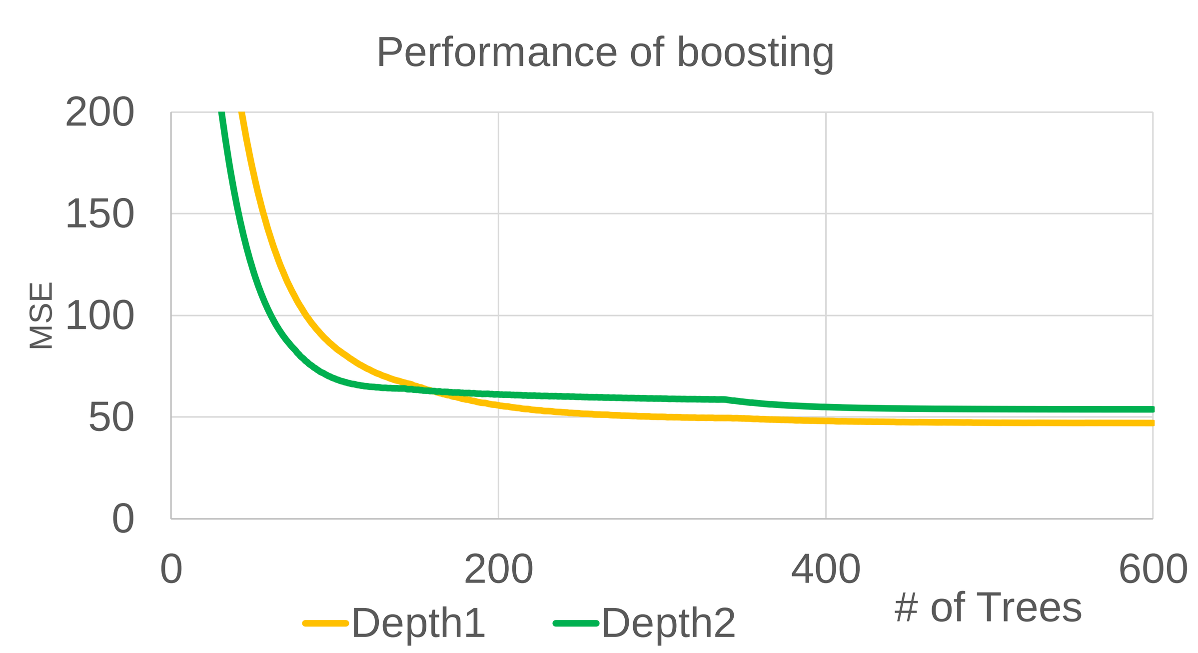
III. Random Forest

An ensemble method that used a weighted sum of multiple random forest model, where each of the random forest models was fit using different parameter settings is used. The weights were inversely proportional to the models' cross-validation error.



IV. Boosting

I also ran a gradient boosting regression tree with shrinking parameter 0.02. Tree depth of 1 and 2 tried.



Results

The simplest model, ensemble sum of linear regression yields the best result, MSE of 36.7. The R^2 of this model is 0.6447. This model is also used to predict the outcome (win/loss) of the test samples based on calculated score estimation, and obtained 88.2% accuracy.

Discussion

I. Passer rate

One of the traditional metric, passer rating has a formula of

$$(rate) = (PCT - 0.3) \times 83.3 + (YPA - 3) \times 4.167$$
$$+ \left(\frac{TD}{ATT}\right) \times 333.3 - \left(\frac{INT}{ATT}\right) \times 416.7$$

A developed linear model using selected features using training data is

$$(score) = (PCT) \times 4.92 + (YPA) \times 2.53$$
$$+ \left(\frac{INT}{ATT}\right)_{opp} \times 55.3 - \left(\frac{INT}{ATT}\right) \times 60.4 + 3.45$$

It turns out traditional passer rating overestimates the importance of pass completion percentage and intercept rate by factor of 10 and 2, compared to average pass yard.

II. First Down, Pass vs Rush

Number of first down is the strongest indicator of team score, but it's not included in traditional metrics or fantasy points. Comparison with average pass and rush yard per attempt showed that pass yards per attempt is much strong indicator to predict the team score, with 3 times bigger coefficient in the linear model.