Predicting Quarterback Fantasy Score

Kevin Thompson, Sean Kennedy, and Sachin Chavan

Data Science Program, Southern Methodist University, USA

"

Abstract

Hype.

1   Introduction

Fantasy sports are a big business. Generating nearly 7 billion dollars annually (cite) and with approximately 60 million players in the US/Canada (cite), fantasy sports – particularly football – have become as big as the sports that they mimic. Players can compete head to head in leagues across a wide range of providers (Yahoo, CBS, ESPN, DraftKings, DraftStreet etc) – each with their own rulesets and stakes. Some are just friendly leagues set up between friends or coworkers, no real stakes other than bragging rights, others have significant monetary rewards for those that can get to the top of the leaderboard.

The explosion of weekly leagues over the course of the last few years has seen an already huge business get even larger. In the weekly cash leagues – each player is given a budget and drafts a completely new team every week. Each draft site uses a predictive model to set player salaries based on the number of points the model predicts that player to score. Budget constraints make it impossible to simply select the players that are predicted to score the most points – hence having a predictive model for which players will generate the best return on investment would be a huge advantage. In the following analysis we will attempt to build position specific models that can accurately predict the number of points a given NFL player is likely to score in a given week.

2   Data Description

3   Exploratory Analysis

4   Objective One

4.1   Model Selection

4.1.1   Type of Selection

Of the tools made available to us in the course, we believe that LASSO estimation is the best option available to us. Subset selection discretely adds and removes variables, leaving far too much room for overfitting.

While ridge regression is a great shrinkage method for many scenarios, we believe that it is at least slightly desirable for the coefficients of truly unnecessary variables to actually be zero so that some true form of variable selection is occurring. We are implicitly saying that a small set of variables are necessary to predict fantasy scores, but the nature of football, the size of our dataset, and the disproportionate amount of points that goes to touchdowns indicates that this assumptions may very well be appropriate.

4.1.2 Assumption-Checking

The LASSO model is a linear model and thus assumes that there is a linear relationship between the response and the parameters. Furthermore, it is also important that the variance of the residuals is constant across the predictors in order for our model to adequately capture the variation in the response. Because we are not interested in quantifying our uncertainty, we don't make any assumptions about the distribution of the residuals.

5 Objective Two