Theory of the Big 3: Predicting NBA Team Win % from Individual Performance

Gene Li, Princeton University '19

Dec. 5, 2016

I. Introduction

Nowhere is the concept of the "Big 3" more relevant than basketball. As a relatively star-dominated game compared to football, soccer, etc., NBA games are determined by the performance of a few players who can deliver offensive firepower. NBA fans often view their team's success as driven by the top three players on each team. Just last season, we saw the trio of Stephen Curry, Klay Thompson, and Draymond Green from the Golden State Warriors face off against Lebron James, Kyrie Irving, and Kevin Love of the Cleveland Cavaliers. Historic "Big Threes" include the infamous James-Wade-Bosh trio in Miami, and the Duncan-Parker-Ginobili Spurs offense that won four championships over 13 years. But just how much can a team's performance be attributed to its top three players?

Fig. 1: San Antonio's "Big Three"

II. Analysis

Is it possible to predict a team's regular season win percentage given player statistics for its top three players? Using historic data from the past 19 years scraped from stats.nba.com, we build a linear regression model that attempts to do so. For every season, we take take each team's top three scorers. We define each players performance vector $\vec{p}$ comprised of features extracted from the stats page.[1]

[1]The feature set is: age, minutes played, field goals made, field goals attempted, 3-pointers made, 3-pointers attempted, free throws made, free throws attempted, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, block attempts, personal fouls, points scored, +/-, double doubles, and triple doubles.

A composite player performance vector is calculated through weighting each of the top three players performance as:

$$\vec{p_c} = w_1\vec{p_1} + w_2\vec{p_2} + w_3\vec{p_3} \tag{1}$$

Optimal weighting w_1, w_2, and w_3 with respect to model accuracy is determined through cross-validation to be $[0.357, 0.5, 0.143]$.

In order to account for differences in statistics across all seasons, for each season we normalize all features each team's $\vec{p_c}$ by standard score calculation.

Thus, we seek a model

$$H(\vec{p_c}) = \text{win }\%. \tag{2}$$

We build this regression model using the popular python machine learning package `scikit-learn`. Through experimentation, we find that simple linear regression model works well. This model is trained on data from the 1997-2013 seasons, and tested on data from 2014-15 and 2015-16 seasons.

III. Results

With the optimal weighting, we generate a linear regression model that predicts regular season win percentages with $R^2 = 0.883$ and an average error $\pm$ 4.25 percentage points. Our model's predictions for the 2014-15 and 2015-16 NBA seasons are plotted in Fig 2.

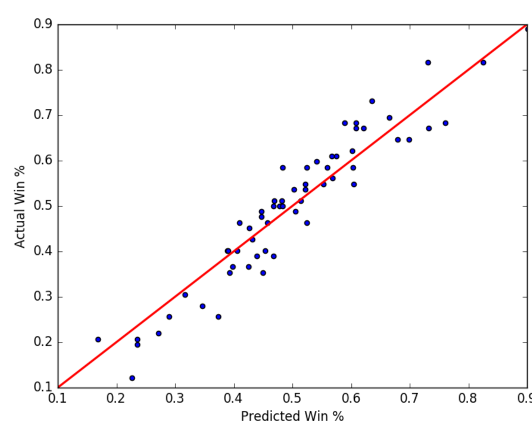

Fig. 2: Predictions for 2014-15 and 2015-16 NBA seasons against Actual Win %

NBA Team	Predicted Win %
GSW	86.9
LAC	84.0
UTA	72.3
CLE	68.2
TOR	65.5
CHI	61.2
OKC	59.4
MEM	57.5
DET	57.4
CHA	54.9
HOU	54.3
SAS	52.8
WAS	48.4
SAC	48.1
BOS	46.2
MIL	45.9
IND	45.6
LAL	44.1
NO	43.2
NYK	41.7
POR	41.2
ATL	40.3
MIA	39.9
DEN	38.1
DAL	37.4
ORL	36.5
MIN	35.8
BKN	33.4
PHI	33.1
PHX	26.4

Fig. 3: Predicted Win Rates for the 2016-17 NBA season

On the whole, this model seems to capture the relative strengths of NBA teams well. However, it does give us some pretty interesting results. For instance, it doesn't seem to like the Spurs much, predicting a 52.8% win rate for a team that is sitting comfortably in second in Western Conference with an 80% win rate. This probably can be explained by Popovich's "team basketball" style, so the top three players by points (Leonard, Aldridge, and Mills) have a reduced workload. It really likes the Jazz, giving them 72.3%, perhaps believing that the trio of Hayward, Hill, and Hood will perform especially well.

IV. Conclusion

A linear regression model proves to be accurate for mapping individual player performance to regular season team win percentages. There are quite a few avenues for future work, including looking matchup specific lineups (given a matchup between two sets of five players with a certain set of performance metrics, who comes out?), as well as measuring how many players on each team actually determine the outcome of games.

For any questions or comments, the author can be reached at gxli@princeton.edu. Source code is available upon request.