An Analysis of Oliver's Four Factors in The Golden Age of NBA Offense

Taaj Cheema

March 18, 2020

1 Introduction

Why are certain NBA teams successful while others are not? There is no shortage of statistics to characterize different aspects of a basketball team's performance, but is there a way we can use just a handful of these to create a model to predict winning, that is both easy to interpret and has strong prediction power?

Introduction needs work, I will come back to it

2 Background

The NBA has been undergoing an offensive renaissance over the past 15 years. We have seen teams prioritize 3-point attempts, improve their effective field goal percentage, and reduce turnovers, to an extent never seen before [3]. This focus on offense has led to record high points per game [13]. There are many factors that have contributed to this offensive boon, but the one that gets the most media attention is the 3-point shot [1] [5] [7] [14].

In his 2004 book, *Basketball on Paper: Rules and Tools for Performance Analysis*, Dean Oliver noted that for basketball, there are four key areas that winning teams excel at [9]. Oliver also notes that for basketball, points per possession, not points per game is the benchmark for offensive efficiency. Oliver asserts that four factors can be used to measure the efficiency of a basketball team's offense or defense. These four offensive factors are effective field goal percentage, turnover percentage, offensive rebound percentage, and free throw factor. The corresponding four defensive factors are the opponent effective field goal percentage, opponent turnover percentage, defensive rebound percentage, and opponent free throw factor [4].

$$eFG\% = \frac{FG + 0.5 * 3FG}{FGA} = \text{effective field goal percentage} \tag{1}$$

$$TOV\% = \frac{TOV}{FGA + 0.44 * FTA + TOV} = \text{turnover percentage} \tag{2}$$

$$OREB\% = \frac{OREB}{OREB + DREB_{OPP}} = \text{offensive rebound percentage} \tag{3}$$

$$FTF = \frac{FT}{FGA} = \text{free throw factor} \tag{4}$$

Each of the four factors has different weights representing its importance to a teams offensive or defensive efficiency. Oliver assigns a weight of 40% to eFG%, 25% to TOV%, 20% to OREB%, and 15% to FTF respectively. The four factors effectively evaluate how good a team is at shooting the ball, how good a team is at protecting the ball, how good a team is at winning offensive rebounds, and how good a team is at shooting free throws.

This research intends to:

1. Use the four factors in linear regression models to predict number of wins and margin of victory
2. Compare the weightings of the different factors to those proposed by Oliver Dean
3. Review how the four factors have historically changed in the NBA

Do you think I should move this to the end of the introduction?

3 Data

All data was obtained from Basketball Reference [2]. The website has an incredible amount of NBA data available.

A dataframe including all NBA teams from 11 NBA seasons spanning from 2008 to 2019 is created using Pandas [10]. The data is only for the regular season, it does not include information on postseason games. The features included in our dataframe are WINS, MOV (5), NRTG (8), eFG% (1), TOV% (2), OREB% (3), FTF (4), OPPeFG%, OPPTOV%, DREB% and OPPFTF. This data is used to create our regression models and to calculate our weightings of the different factors.

$$MOV = \frac{TotPTS - TotPTS_{OPP}}{82} = \text{margin of victory} \tag{5}$$

$$ORTG = 100 * \frac{TotPTS}{TotPoss} = \text{offensive rating} \tag{6}$$

$$DRTG = 100 * \frac{TotPTS_{OPP}}{TotPoss_{OPP}} = \text{defensive rating} \tag{7}$$

$$NRTG = ORTG - DRTG = \text{net rating} \tag{8}$$

A dataframe is also created on NBA league averages per 100 possessions for 20 NBA seasons spanning from 2000-2000. This data is used to visualize historical changes in the four factors.

4 Methods

The data needs to undergo certain preprocessing tasks before we can use it for modeling.

Certain features in the dataframe are scaled. The eFG%, FTF, OPPeFG% and OPPFTF features are multiplied by 100. This is done because in the original dataframe, these features are in decimal form, whereas the other features that will be used as predictors for the model are in percentage form. Additionally, the 2011-2012 NBA season was a lockout season, so teams only played a total of 66 games. The wins feature for records in the 2011-2012 season was scaled by (82/66).

The data is then normalized using the preprocessing.normalize function in scikit-learn [11]. This is done because we want to calculate wins using normalized change in each feature. This is done with the L2 norm.

This data is then split into random train and test subsets. 70% of the original dataset is randomly partitioned to x_train and y_wins_train, while the other 30% is partitioned to x_test and y_wins_test. This is done using the train_test_split function included in scikit-learn [11]. A random state variable is set to create a reproducible output.

Two linear regression models using the scikit-learn package [11]. The ordinary least squares method is used to fit the regression. The OLS method fits a linear model with coefficients beta 0, beta 1, beta 2 to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation [6].

The first model fits the four factors as features to predict wins.

$$\begin{aligned} WINS = \beta_0 + \beta_1 * eFG\% + \beta_2 * TOV\% + \beta_3 * OREB\% \\ + \beta_4 * FTF + \beta_5 * OPPeFG\% + \beta_6 * OPPTOV\% \\ + \beta_7 * DREB\% + \beta_8 * OPPFTF + \epsilon \end{aligned} \tag{9}$$

The second model fits the four factors as features to margin of victory.

$$\begin{aligned} MOV = \beta_0 + \beta_1 * eFG\% + \beta_2 * TOV\% + \beta_3 * OREB\% \\ + \beta_4 * FTF + \beta_5 * OPPeFG\% + \beta_6 * OPPTOV\% \\ + \beta_7 * DREB\% + \beta_8 * OPPFTF + \epsilon \end{aligned} \tag{10}$$

After training our model we get our fitted equations

$$\begin{aligned} WINS = & -568.15 + 680.22 * eFG\% - 407.49 * TOV\% + 234.38 * OREB\% \\ & + 195.95 * FTF - 190.75 * OPPeFG\% + 348.65 * OPPTOV\% \\ & + 456.11 * DREB\% + 13.08 * OPPFTF + \epsilon \end{aligned} \tag{11}$$

$$\begin{aligned} MOV = \beta_0 + \beta_1 * eFG\% + \beta_2 * TOV\% + \beta_3 * OREB\% \\ + \beta_4 * FTF + \beta_5 * OPPeFG\% + \beta_6 * OPPTOV\% \\ + \beta_7 * DREB\% + \beta_8 * OPPFTF + \epsilon \end{aligned} \tag{12}$$

I'm going to come back to this and update it with the coefficients from the OLS model obtained with the statsmodel package, this is in the modeling Jupyter notebook

The test data can now be applied to the model and we can output our predicted wins.

5 Results

The model produces statistically significant p-values for all features. The model was successful in predicting NBA wins and achieves a R-squared value of 0.9262. The model explains 92.62% of the variation in wins.

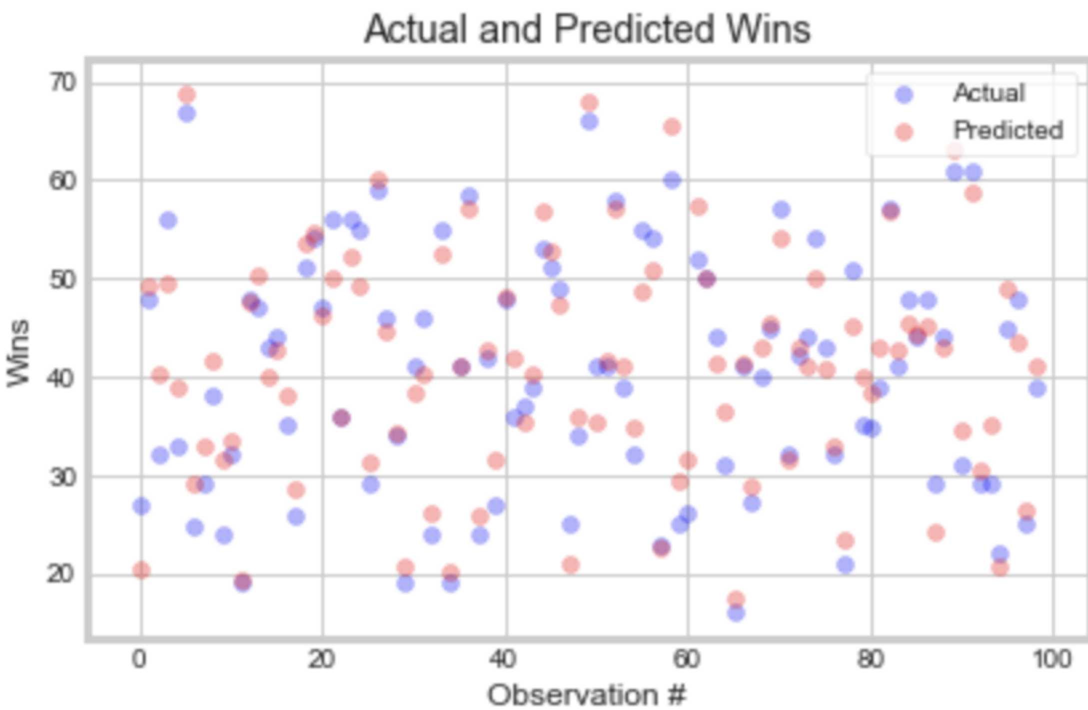

Figure 1: Actual Wins vs. Predicted Wins

Effective Field Goal percentage has been trending upwards in the NBA over the past twenty years. Turnover percentage, offensive rebound percentage and free throw factor have all been trending downward.

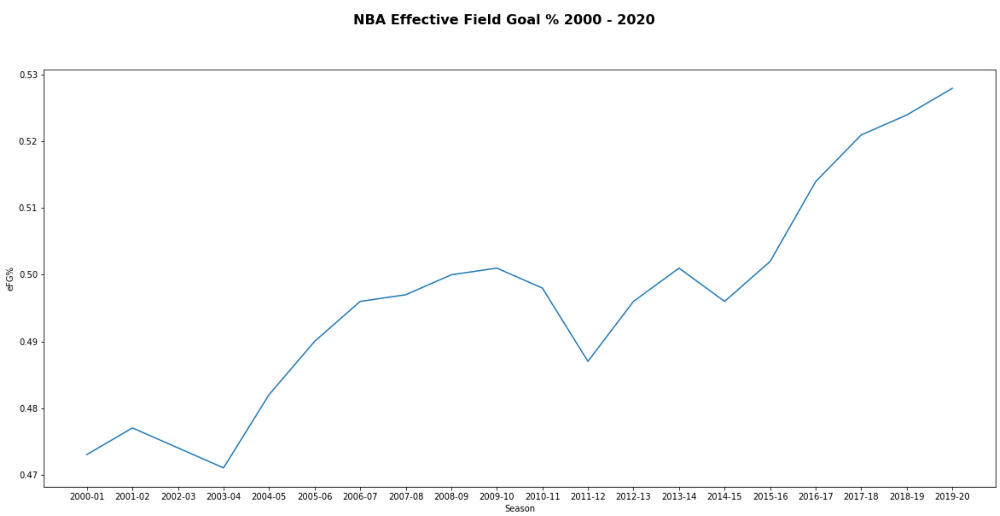

Figure 2: NBA Effective Field Goal % from 2000 to 2020

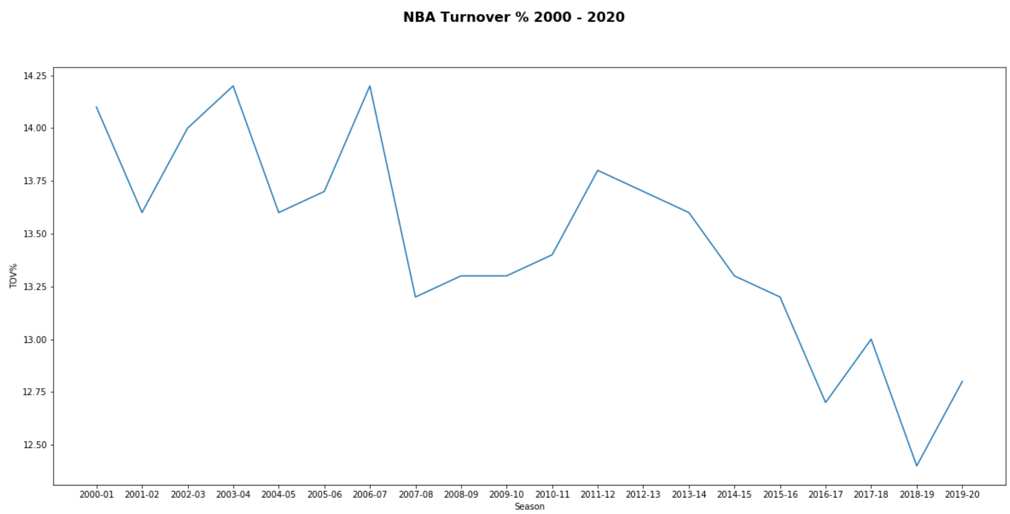

Figure 3: NBA Turnover % from 2000 to 2020

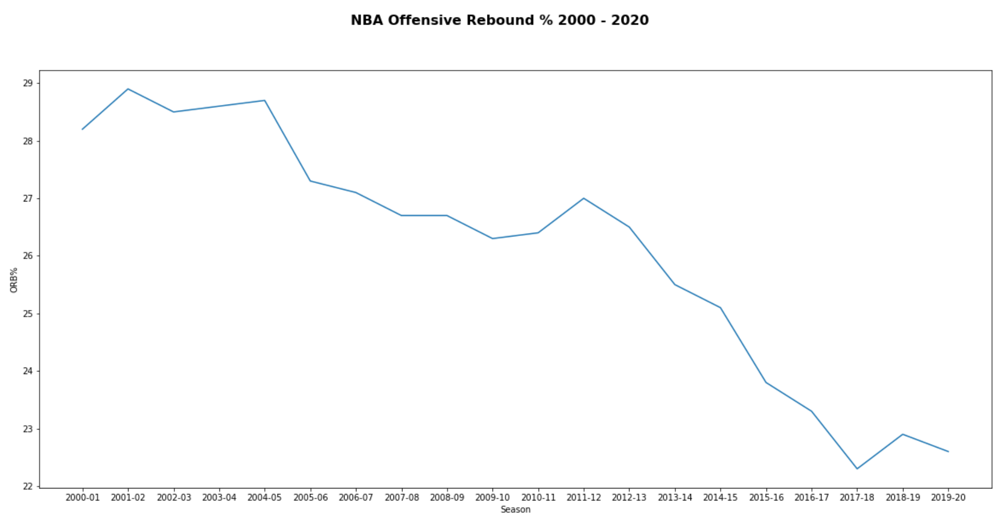

Figure 4: NBA Offensive Rebound % from 2000 to 2020

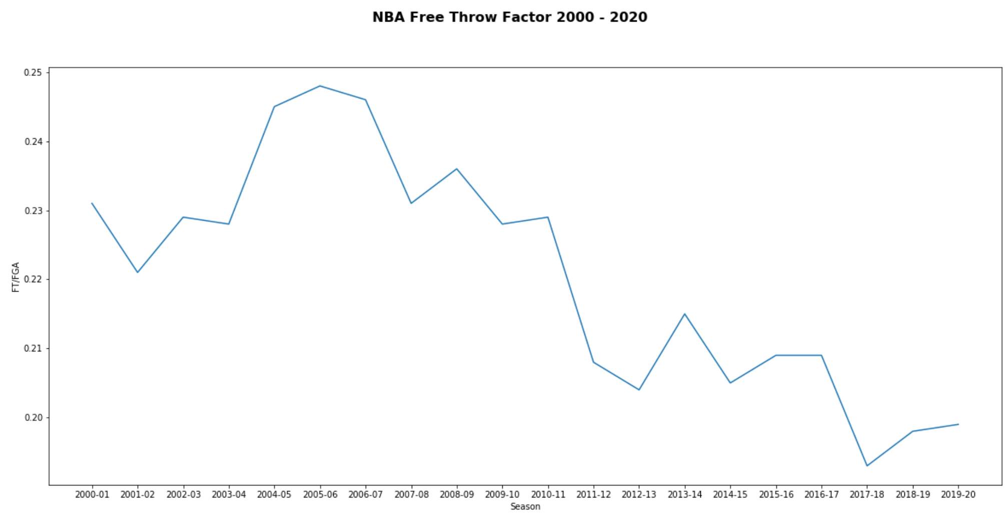

Figure 5: NBA Free Throw Factor from 2000 to 2020

6 Conclusion

A multiple linear regression model to predict NBA team wins using Dean's Four Factors was successfully created. The Four Factors provide us an effective way to represent a team's overall performance. The weightings we obtained for the different factors were similar to those proposed by Dean. I would like to try more models. In the future try a gradient boosting model and a random forest model.

References

[1] Cohen, Ben. "The First Shots of the NBA's 3-Point Revolution". *The Wall Street Journal*, 12 Mar. 2018, `www.wsj.com/articles/the-first-shots-of-the-nbas-3-point-revolution-1523542076`.

[2] "Basketball Reference". *Basketball Reference*. `www.basketball-reference.com`.

[3] Dubin, Jared. "The NBA's Other Offensive Revolution: Never Turning The Ball Over". *FiveThirtyEight*, 14 Mar. 2019, `www.fivethirtyeight.com/features/the-nbas-other-offensive-revolution-never-turning-the-ball-over/`.

[4] "Four Factors". *Basketball Reference*. `www.basketball-reference.com/about/factors.html`.

[5] "Gametime: 3-Point Revolution Part One". *NBA*.

[6] James, Gareth, et al. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2017.

[7] Kram, Zach. "The 3-Point Boom Is Far From Over". *The Ringer*, 27 Feb. 2019, `www.theringer.com/nba/2019/2/27/18240583/3-point-boom-nba-daryl-morey`.

[8] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan. (2007) "A Starting Point for Analyzing Basketball Statistics". *Journal of Quantitative Analysis in Sports*. Volume 3: Issue 3, Article 1.

[9] Oliver, Dean. *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, 2004.

[10] "Pandas API Reference". *Pandas*. `www.pandas.pydata.org/docs/reference/index.html`.

[11] "Scikit-learn API Reference". *Scikit-learn*.
`www.scikit-learn.org/stable/modules/classes.html`.

[12] Winston, Wayne. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*. Princeton University Press, 2012.

[13] Woike, Dan. "107 Is the New 100.' How Scoring Trends in the NBA Are Setting a New Bar for What a Good Defense Is.". *Los Angeles Times*, 29 Oct. 2018,
`www.https://www.latimes.com./sports/nba/la-sp-nba-defenses-20181029-story.html`.

[14] Young, Shane. "The NBA's 3-Point Revolution Continues To Take Over.". *Forbes*, 1 Dec. 2019,