Predicting Team Success in the NBA using Oliver's Four Factors

Taaj Cheema

1   Introduction

How can we distill the success of a basketball team? Can the performance of a team be modeled by just a few key components? There is no shortage of box score and advanced statistics to characterize different aspects of a team's performance. Ideally, we would like to use a small selection of readily available and understandable statistics that can capture most of the variation in a given target metric of success. Furthermore, can we can use these features to predict a team's wins and average margin of victory using historical National Basketball Association data?

In this paper we will test if the four factors posited by Dean Oliver have statistically significant prediction power on the success of an NBA team. Additionally, we will use various machine learning models to predict the success of basketball teams in the NBA using the four factors as features. We define success with two output variables, number of wins in a season and average margin of victory. Lastly, we will calculate the respective weights of these four factors in predicting success and compare them to the weights proposed by Oliver and Ed Küpfer.

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 (Woike 2018). This focus on offense has led to record high points per game (Dubin 2019). 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 (NBA 2017); (Cohen 2018); (Young 2019).

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 (Oliver 2004). Oliver also notes that for basketball, points per possession, not points per game is the benchmark for offensive efficiency. Oliver asserts that the four factors encapsulate effective basketball possessions and can be used to measure the efficiency of a basketball team's offense or defense. The four offensive factors are effective field goal percentage (eFG%), turnover percentage (TOV%), offensive rebound percentage (OREB%), and free throw factor (FTF). The corresponding four defensive factors are opponent effective field goal percentage, opponent turnover percentage, defensive rebound percentage, and opponent free throw factor (BasketballReference 2018).

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

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

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

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

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 (Oliver 2004). In later analysis by Houston Rockets analyst Ed Küpfer, Küpfer assigns a weight of 10, 6, 3, and 3 to eFG%, TOV%, OREB%, and FTF respectively (Kubatko et al. 2007). This works out to approximately 45.45%, 27.27%, 13.64% and 13.64%.

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 rebounds, and how good a team is at getting to the free throw line and successfully converting these attempts. In theory, the four factors should be able to predict the success of a NBA team.

This research intends to:

1. Test if the four factors are statistically significant predictors of success in the NBA

2. Use the four factors to predict a team's number of wins and average margin of victory in the NBA using historical data and various machine learning models

3. Compare the observed weightings of the different factors in our models to those proposed by Dean Oliver and Ed Küpfer.

All of the code, the write up, as well as additional resources related to this paper are available at:
https://github.com/taajcheema/four_factors

3   Related Work

Similar research to model NBA games has been performed before.

In a 2003 paper, (Hoffman and Joseph 2003), by Lori Hoffman and Maria Joseph, the researchers predicted if teams would qualify for the NBA playoffs using methods including principal components analysis and discriminant analysis. They identified 5 principal components that were used to accurately predict 26 out of 29 teams as either playoff or non-playoff teams for the 2002-2003 NBA season (Hoffman and Joseph 2003).

A 2008 paper by Beckler, Papamichael, and Wang titled *NBA Oracle* sought to predict the winner of individual NBA games from the 1991-1992 through 1996-1997 seasons using linear regression, logistic regression, support vector machines, and neural networks. They were able to correctly predict the winner 73% of the time using linear regression and a total of 60 features. This is compared to a baseline of 50% by choosing a winner randomly (Beckler, Wang, and Papamichael 2008).

A 2013 paper by Renato Torres, *Prediction of NBA games based on Machine Learning Methods*, also predicted the winner of individual NBA matchups using statistical methods. He was able to correctly predict the winner of 68.44% of NBA games in the 2006-2007 to 2011-2012 seasons using a multilayer neural network and 8 features (Torres 2013).

A 2014 paper by Lin, Short, and Sundaresan, *Predicting National Basketball Association Winners*, predicted the winner of NBA games using various machine learning methods. They used 16 independent variables to predict the winner. The researchers found that a basketball team's win record plays a central role in determining their likeliness of winning future games. When they removed the win record feature from their

model, the classification accuracy decreased significantly. They concluded that traditional box score statistics fail to represent all of a team's success on the court (Lin, Short, and Sundaresan 2014).

A 2016 paper by Avalon, Balci, and Guzman *Various Machine Learning Approaches to Predicting NBA Score Margins*, predicted the margin of victory of two NBA teams using linear regression, Gaussian discriminant analysis, principal component analysis coupled with support vector machines, random forest and adaptive boosting. They started with 218 features that they reduced into the most important components using PCA. They trained their models on 1052 games and tested them on 264 games all from the 2013 - 2014 NBA seasons. With a PCA and SVM model, they were able to predict the margin of victory within a 2 point score margin of error for 90% of the teams in the test set (Avalon, Balci, and Guzman 2016).

Our paper differs from the related works in that we are attempting to predict the number of wins and average margin of victory for teams over a season, rather than predicting the winner in individual matchups. We are additionally testing the significance of the four factors of basketball success proposed by Dean Oliver, and using them as features in our models. Lastly, we are assigning a weighting to each of the factors using our models.

4   Data

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

A dataframe is created including statistics for all NBA teams from 11 seasons spanning from the 2008-2009 season to the 2018-2019 season. 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 predictive models and to calculate our weightings of the four factors.

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

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

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

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

A dataframe is also created on NBA league averages per 100 possessions for 20 NBA seasons spanning the 2000-2001 to the 2019-2020 season. This data is used to visualize historical changes in the four factors.

5   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. Furthermore, 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 is scaled by (82/66). Additionally, some of column names in the dataframe are changed to make them easier to work with.

Once our data is prepared, we analyze it graphically with ggplot2 by creating a variety of graphs. This is done to check the basic assumptions of mulitple linear regression. We create scatterplots of the predictor and response variables to see if there is a relationship between them. We also create boxplots to check for outliers, along with density plots to make sure the features follow a normal distribution. We additionally use the Durbin-Watson test to check for potential autocorrelation in the residuals of our regression models. Finally, we use pairplots and correlation plots to check for multicollinearity.
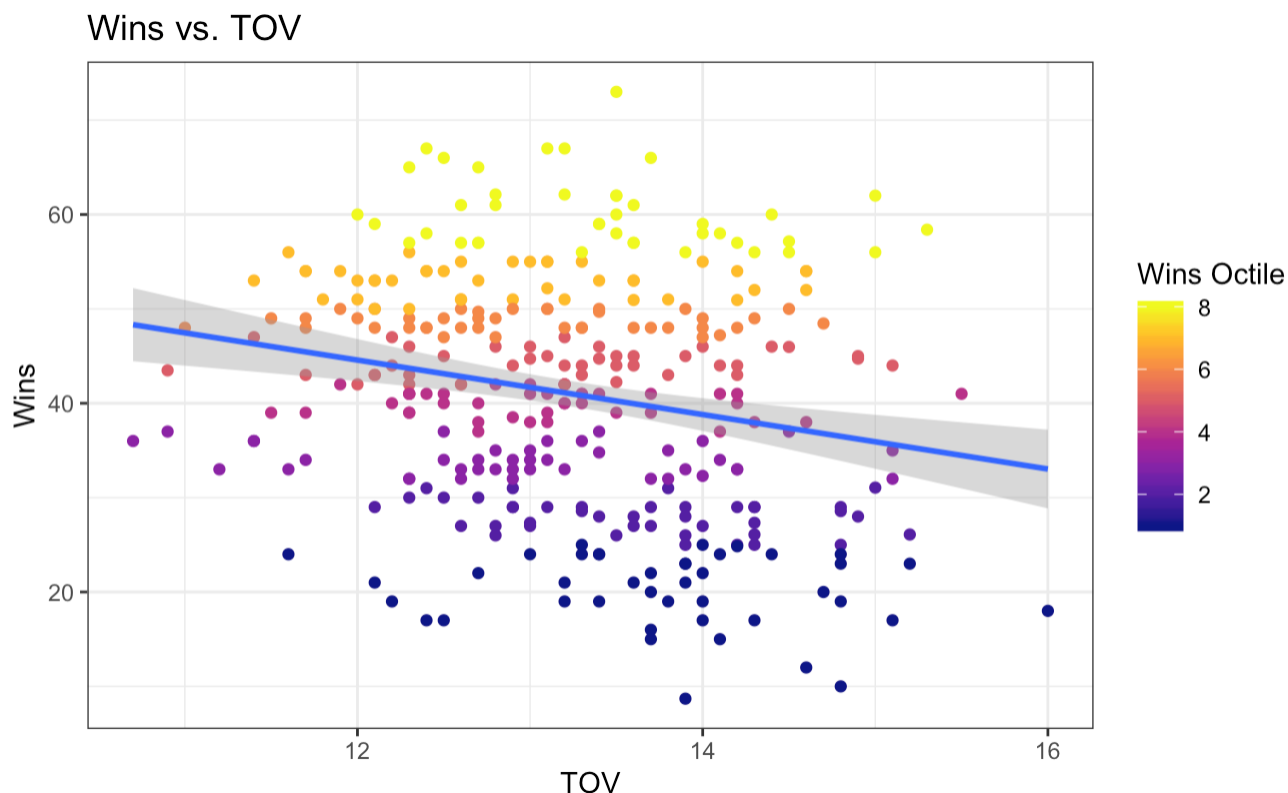

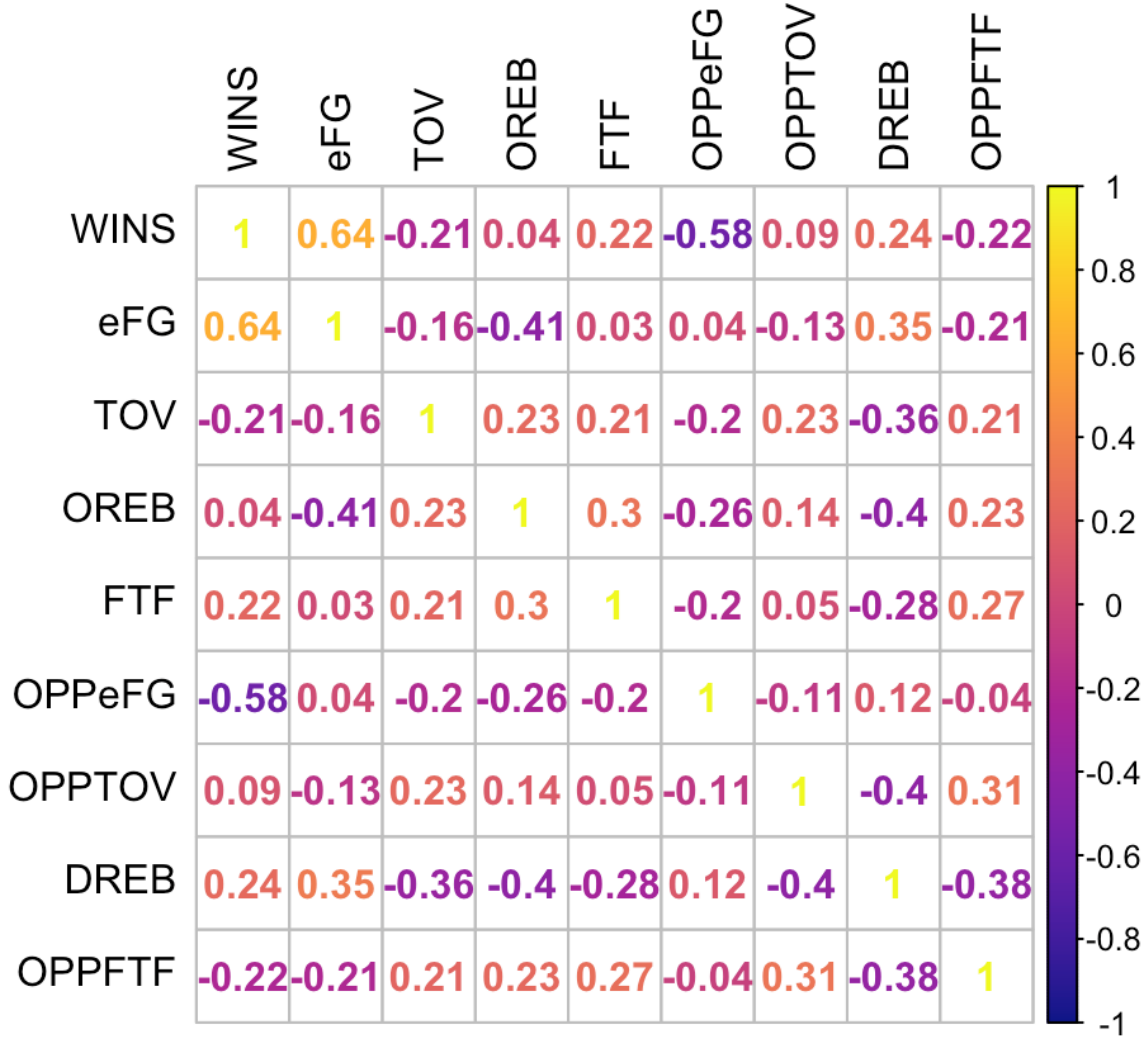**Figure 1:** Scatter plot of WINS vs TOV%

Figure 2: Correlation matrix

This data is then split into random train and test subsets. 70% of the original dataset is randomly partitioned to ff_train while the other 30% is partitioned to ff_test. This is done using the sample function in R. The set.seed() function is used to create a reproducible output.

Two linear regression models are fit using the lm() function in R using the ordinary least squares method. The OLS method fits a linear model and minimizes the residual sum of squares between the observed and predicted targets in the dataset (Gareth et al. 2014).

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

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

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

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

After training our model we get our fitted equations

$$WINS = -53.185 + 3.899 \times eFG\% - 3.392 \times TOV\% + 1.111 \times OREB\%$$
$$+ 0.708 \times FTF - 3.765 \times OPPeFG\% + 2.902 \times OPPTOV\%$$
$$+ 0.887 \times DREB\% - 0.729 \times OPPFTF + \epsilon \tag{11}$$

$$MOV = -38.851 + 1.436 \times eFG\% - 1.323 \times TOV\% + 0.438 \times OREB\%$$
$$+ 0.296 \times FTF - 1.420 \times OPPeFG\% + 1.170 \times OPPTOV\%$$
$$+ 0.386 \times DREB\% - 0.288 \times OPPFTF + \epsilon \tag{12}$$

Regression results

	WINS (1)	MOV (2)
eFG	3.899*** (0.110)	1.436*** (0.024)
TOV	-3.392*** (0.249)	-1.323*** (0.054)
OREB	1.111*** (0.090)	0.438*** (0.020)
FTF	0.708*** (0.093)	0.296*** (0.020)
OPPeFG	-3.765*** (0.114)	-1.420*** (0.025)
OPPTOV	2.902*** (0.212)	1.170*** (0.046)
DREB	0.887*** (0.118)	0.386*** (0.026)
OPPFTF	-0.729*** (0.104)	-0.288*** (0.023)
Constant	-53.185*** (14.548)	-38.851*** (3.165)
N	230	230
R^2	0.943	0.981
Adjusted R^2	0.941	0.980
Residual Std. Error (df = 221)	3.158	0.687
F Statistic (df = 8; 221)	461.199***	1,398.216***

*$p < .1$; **$p < .05$; ***$p < .01$

Figure 3: Regression results of the lmWINS and lmMOV models

We also performed diagnostic tests to make sure that we do not violate any assumptions of multiple linear regression. Using the plot() function in R, we created plots for residuals vs. fitted, normal Q-Q, scale-location, and Cook's distance. We also checked the variance inflation factor of our models to make sure there is no multicollinearity in our features. The results of the diagnostic tests can be seen in figures 5 and 6.
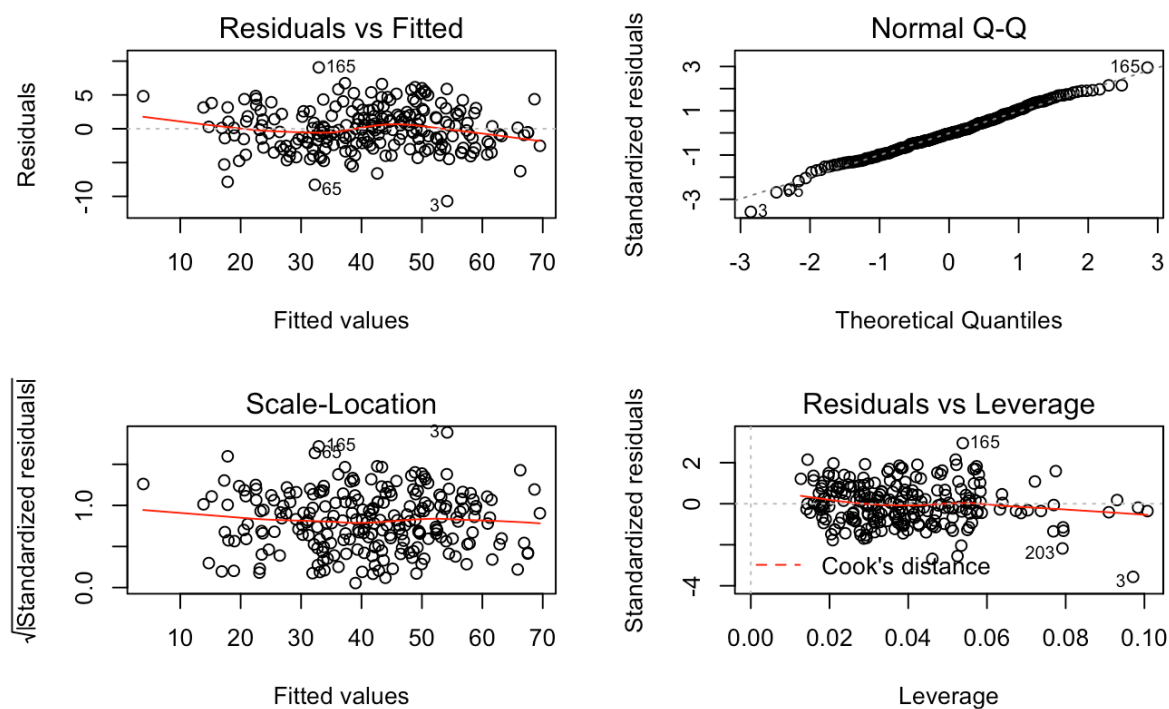

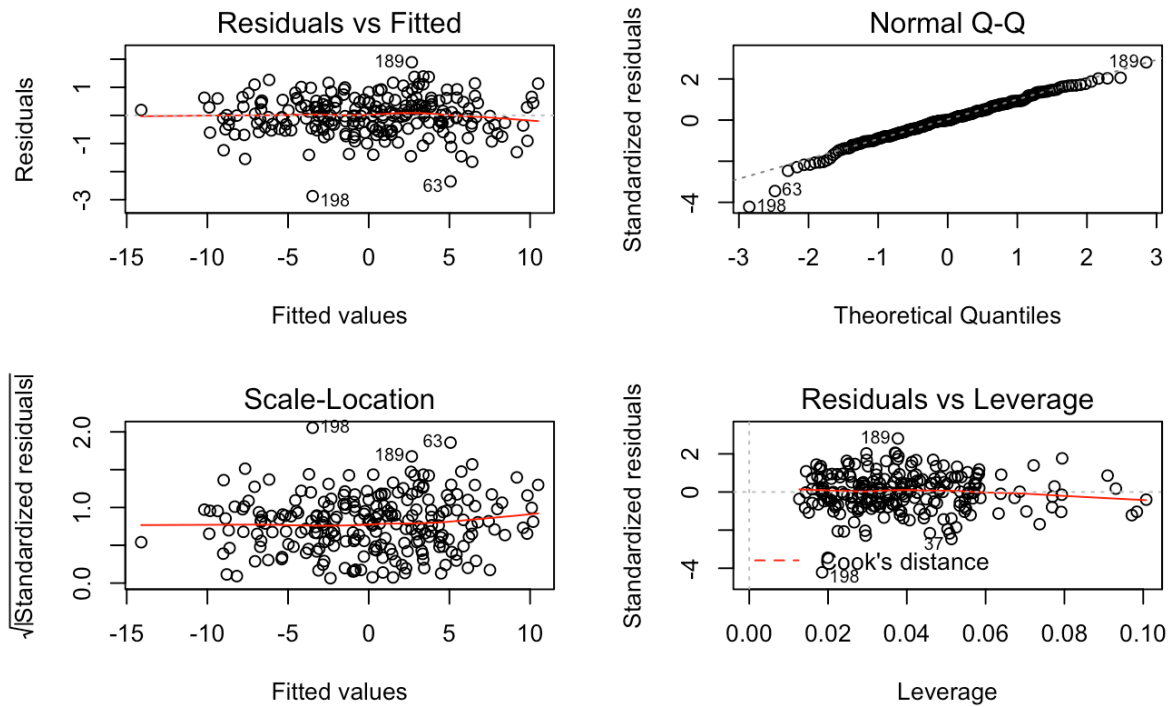**Figure 5:** Diagnostic plots for the lmWINS model

Figure 6: Diagnostic plots for the lmMOV model

We additionally made a random forest model and a gradient boosting machine model to predict Wins and margin of victory.

6   Results

Wins and margin of victory were predicted using three different models

1. Multiple Linear Regression

2. Random Forest

3. Gradient Boosting Machine

For the final model, we decided to use multiple linear regression. The linear regression models we fit performed better than the random forest and gradient boosting machine models. Had we further tweaked and tested different hyperparameter configurations for the other models, they may have been able to perform better than the linear regression model.

Another reason we picked multiple linear regression as the final model type, is because it is the simplest model. Simple models are often easier to explain. In this case, the other models are black-box models, which can potentially provide more accuracy, but at the expense of explainability compared to a model like multiple linear regression.

6.1   Are the four factors statistically significant predictors of success in the NBA?

The p-values for both the lmWINS and lmMOV models are practically zero. This is strong evidence that we can reject the null hypothesis that there is no relationship between our response and predictor variables.

Additionally, both the lmWINS and lmMOV models produce statistically significant p-values for all included features at a 0.001 significance level. All of the included predictors have statistically significant value in predicting wins and margin of victory respectively.

From the plots in figures 5 and 6 we can see that we reasonably meet the assumptions of linear regression. The residual plot is mostly flat for both the lmWINS and lmMOV models. There are a few points skew the line, but overall the relationship is linear. In the scale-location plot, points are normally spread. There is no evidence of potential heteroskedasticity. In the Q-Q plot, points are very close to the diagonal line. We can assume that values are normally distributed. From the Cook's Distance plot, we can see there are a few outliers. We choose to leave these points in because we know that they were real data, and not errors. We also checked the variance inflation factor to make sure there is no multicollinearity in our features. All of the VIF's for our features were below 2 and we can conclude that there is no multicollinearity in our data.

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.

6.2   Model Observed vs. Predicted

The lmWINS model was successful in predicting NBA wins and achieves an adjusted R-squared value of 0.9435. The can be interpreted as our model is able to explain 94.35% of the variation in wins. When we fit the lmWINS model on the test data, we achieve a an adjusted R-squared value of 0.9300. This shows that our lmWINS model generalizes well. The linear model fit well and could potentially be used to predict future wins.

The lmMOV model achieved a R-squared value of 0.9785. The can be interpreted as our model is able to explain 97.85% of the variation in margin of victory. When we fit the lmMOV model to the test data, we achieve a R-squared value of 0.9707. The model again generalizes well on the test data.

The results were very good. The four factors proposed by Oliver can be highly effective in predicting wins and margin of victory in the NBA.

Team	Season	Observed Wins	Predicted Wins	Observed Margin of Victory	Predicted Margin of Victory
Milwaukee Bucks	2018 - 2019	60	61.01	8.87	7.50
Denver Nuggets	2018 - 2019	54	48.44	3.95	2.79
Boston Celtics	2018 - 2019	49	51.48	4.44	3.86
Brooklyn Nets	2018 - 2019	42	41.65	0.71	0.13
Charlotte Hornets	2018 - 2019	39	38.43	-0.23	-0.86
Washington Wizards	2018 - 2019	32	35.06	-2.90	-2.42
Cleveland Cavaliers	2018 - 2019	19	17.25	-9.34	-9.02
Phoenix Suns	2018 - 2019	19	18.57	-9.61	-8.79
Houston Rockets*	2017 - 2018	65	59.86	8.48	7.26
Toronto Raptors*	2017 - 2018	59	58.06	7.78	6.38
Golden State Warriors*	2017 - 2018	58	58.90	5.98	6.39
Portland Trail Blazers*	2017 - 2018	49	44.47	2.60	1.33
Indiana Pacers*	2017 - 2018	48	44.95	1.38	1.46
San Antonio Spurs*	2017 - 2018	47	47.56	2.89	2.63
Milwaukee Bucks*	2017 - 2018	44	42.03	-0.30	0.33

Figure 7: Observed vs. predicted values for wins and margin of victory for the first 15 observations in our test data

6.3 Weightings

We obtained weights for the different features of the lmWINS and lmMOV models by averaging the absolute value of each of the coefficients. An example where the relative weight of the eFG% factor for the lmWINS model is derived can be seen in equations 13, 14 and 15.

$$
\begin{aligned}
sum\_all\_lmWINS = |3.89934| + |-3.39168| + |1.11070| + \\
|0.70791| + |-3.76483| + |2.90168| + \\
|0.88666| + |-0.72897|
\end{aligned}
\tag{13}
$$

$$sum\_lmWINS = |3.89934| + | - 3.76483| \tag{14}$$

$$weight\_eFG\_lmWINS = 100 \times (sum\_eFG\_lmWINS/sum\_all\_lmWINS) \tag{15}$$

The order of importance of the weights that we obtained with our model agree with both Oliver and Küpfer. We found shooting or eFG% to be the most important of the four factors to a team's success. Turnovers, rebounding and free throw factor followed respectively. However, the weights we obtained for the four factors are different. We created our model on regular season NBA games from the 2008 - 2009 season to the 2018 - 2019 season. Oliver originally created his model in 2002 and Küpfer in 2005, so it is possible the importance of the weights has shifted as the NBA has changed.

Factor	Oliver	Küpfer	lmWINS	lmMOV
Shooting	0.40	0.4545	0.4407	0.4226
Turnovers	0.25	0.2727	0.3619	0.3691
Rebounding	0.20	0.1364	0.1148	0.1219
Free Throw Factor	0.15	0.1364	0.0826	0.0865

Figure 8: Comparison of weights for the four factors

7 Conclusion

A multiple linear regression model to forecast NBA teams' wins and margins of victory using Oliver's Four Factors was successfully created. The Four Factors provide us an effective way to represent a team's overall performance in a relatively simple and easy to understand model.

The weightings we obtained for the four factors were similar to those proposed by Dean, with our model prioritizing turnovers and deprioritizing rebounding and free throw factor.

A future work will focus on predicting winners and margin of victory of individual NBA matchups, rather than overall season predictions, using a similar four factors model. This would entail training a model on 1230 games per NBA season, rather than the 30 data points of summary statistics per NBA season we trained the models for this paper on. Another future approach will involve looking at only NBA playoff games. It would be interesting to see how well the four factors can predict success in the NBA postseason as it is often much more competitive and high stakes than the regular season. One final future approach will use a binary logistic regression model to predict probabilities of two opposing teams winning in matchups, using the four factors as features.

References

Avalon, G, B Balci, and J Guzman. 2016. "Various Machine Learning Approaches to Predicting NBA Score Margins," http://cs229.stanford.edu/proj2016/report/Avalon_balci_guzman_various_ml_approaches_NBA_Scores_report.pdf.

BasketballReference. 2018. "Four Factors," https://www.basketball-reference.com/about/factors.html.

Beckler, Matthew, Hongfei Wang, and Michael Papamichael. 2008. "NBA Oracle," http://www.cs.cmu.edu/~mpapamic/projects/NBA_Oracle_midway_report.pdf.

Cohen, Ben. 2018. "The First Shots of the NBA's 3-Point Revolution." *The Wall Street Journal* (April 12, 2018). https://www.wsj.com/articles/the-first-shots-of-the-nbas-3-point-revolution-1523542076.

Dubin, Jared. 2019. "The NBA's Other Offensive Revolution: Never Turning The Ball Over." *FiveThirtyEight* (March 14, 2019). https://fivethirtyeight.com/features/the-nbas-other-offensive-revolution-never-turning-the-ball-over/.

Gareth, J., D. Witten, T. Hastie, and R. Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated. ISBN: 1461471370.

Hoffman, Lori, and Maria Joseph. 2003. "A Multivariate Statistical Analysis of the NBA," http://www.units.miamioh.edu/sumsri/sumj/2003/NBAstats.pdf.

Kubatko, Justin, Dean Oliver, Kevin Pelton, and Dan T Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports* 3 (3). https://statsbylopez.files.wordpress.com/2016/01/jqas-2007-3-3-1070.pdf.

Lin, Jasper, Logan Short, and Vishnu Sundaresan. 2014. "Predicting National Basketball Association Winners," https://www.semanticscholar.org/paper/Predicting-National-Basketball-Association-Winners-Lin-Short/d6ec5a118d7d4f497b22610a5472574aa821fc9f.

NBA. 2017. "GameTime: The Growth of the 3-pointer." *NBA,* https://www.nba.com/amp/league/video/2017/12/04/20171204-3point-revolution-episode-3-analysis.

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

Torres, Renato Amorim. 2013. "Prediction of NBA games based on Machine Learning Methods," https://homepages.cae.wisc.edu/~ece539/fall13/project/AmorimTorres_rpt.pdf.

Woike, Dan. 2018. "'107 is the new 100.' How scoring trends in the NBA are setting a new bar for what a good defense is." *LA Times* (October 29, 2018). https://www.latimes.com/sports/nba/la-sp-nba-defenses-20181029-story.html.

Young, Shane. 2019. "The NBA's 3-Point Revolution Continues To Take Over." *Forbes* (November 30, 2019). https://www.forbes.com/sites/shaneyoung/2019/11/30/the-nbas-three-point-revolution-continues-to-take-over/?sh=2eee9e8656b8.