Swaraj Rai

QAC211

Professor Maryam Gooyabadi

05/16/2023

Re-examining Dean Oliver's determinants of winningness in the NBA

Abstract: The fluidity of play styles in NBA basketball calls for a reassessment of prominent research conducted by Dean Oliver on the statistical determinants of winning in NBA basketball due to its publication back in 2004. How statistically significant are Dean Oliver's factors of winningness in the modern NBA era and are there any potential unencountered factors that are worth considering in understanding the impact of winning-ness in the NBA? This study takes data from the 2010/11 season to the 2020/21 season and runs all sorts of multiple linear regressions, correlation tests, PCA, clustering, and cross-validation to assess winningness in modern NBA basketball. The best model has the highest adjusted R^2 and lowest AIC excluding one of the 4 original factors, FT%, and added 3P% and STL. These findings conclude that Oliver's publications are indeed outdated and don't include key statistics like 3P% or STL, which create regressions with the highest explained variation and lowest AIC values.

Background: This project looks to examine statistical significance in the evolving playstyle of NBA basketball. The research question of this paper is: How statistically significant are Dean Oliver's factors of winning-ness in the modern NBA era and are there any potential unencountered factors that are also worth considering in understanding the impact of winning-ness in the NBA? First, it's important to define winning-ness in this context, which is essentially a team's TotalWins/Total Wins+Losses. Secondly, when looking at the prior literature available on the subject, it becomes apparent that understanding statistical determinants of winning in basketball has long been researched before Dean Oliver's publications, some as far back as nine years before Oliver's release[1], however, his have widely been regarded and referenced to since its release in 2004. Secondly, identifying that there is any basis for asserting that there could be enough of a change in the play style of basketball to warrant such research. To make such claims, I use literature examining the statistical data of NBA and Euroleague basketball and their trends to determine the fact that not only are statistical drivers of types of shots (2P/3P) in basketball appear to be shifting, the Euroleague and NBA appear to be converging in playstyles[2].

The NBA is the most competitive basketball league in the world, and it makes sense why. The game originated in the United States in the 1930s and since has been the pinnacle of basketball. However, as the game of basketball has spread like wildfire across the globe, play styles have fragmented and made for interesting studies on common trends and methods of play that remain common amongst the world's best. The hot topic of discussion in this current era of the NBA is the 3-pointer: the highest point shot (aside from a 4-point play) in which the attempts made of making such a shot have gone up by considerable numbers in the past 13 years alone. Over the last 40 years, research has indeed confirmed that the ratio of 3-point/2-point shots has shifted dramatically[3], where efficiencies in both categories have either gone up or remained

[1] Chatterjee, S., Campbell, M. R., & Wiseman, F. (1994). *Take that jam! an analysis of winning percentage for NBA teams*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/mde.4090150514

[2] Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019, October 7). *Trends in NBA and Euroleague Basketball: Analysis and comparison of statistical data from 2000 to 2017*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0223524

[3] Zając, T., Mikołajec, K., Chmura, P., Konefał, M., Krzysztofik, M., & Makar, P. (2023). Long-Term Trends in Shooting Performance in the NBA: An Analysis of Two- and Three-Point Shooting across 40 Consecutive Seasons.

stable in the timeframe. This represents an overall increase in efficiencies in NBA basketball and can be attributed to the players of today's game being far more skilled than 40 years ago, which is a separate study on its own. This study, however, focuses on the evolution of the game in conjunction with Dean Oliver's Four Factors of "Winningness" in the NBA and whether such factors pose statistical significance in the period of the NBA we are in by testing NBA team data of the past 10 years on the same factors Dean Oliver considered to be vital to the success of an NBA team. These statistics are Field Goal Percentage (referred to as FG%), which is just shots made/shots attempted (excluding free throws). The second statistic is the Total Turnover percentage (referred to as TOV%) (Jacobs)[4], which is measuring how often a team turns over the ball (Turnover amount/Number of possessions in a game). Third is Offensive Rebound Percentage, how often a team can secure the ball on offense after missing a shot. The last statistic is the free throw rate, which measures the ability of a team to capitalize on getting to the free throw line (total free throws makes/total free throws attempted). By the end of this project, I'm hoping to achieve some clarity on how this era may differ from that of the past (pre-2010s) and where points of emphasis in the NBA play-style are placed.

This project seeks to incorporate past findings of statistical significance on outdated data that no longer remains relevant and seeks to inquire whether such findings remains relevant. Basic regression analysis with different models utilizing different variables (team statistics) in basketball on data deemed relevant to the team to provide insight into the current nature of basketball in the NBA.

International journal of environmental research and public health, *20*(3), 1924. https://doi.org/10.3390/ijerph20031924

[4] Jacobs, Justin. "Introduction to Oliver's Four Factors." *Squared Statistics: Understanding Basketball Analytics*, squared2020, 5 September 2017, https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/. Accessed 11 April 2023.

Methods: *How statistically significant are Dean Oliver's factors of winning-ness in the modern NBA era and are there any potential unencountered factors that are also worth considering in understanding the impact of winning-ness in the NBA?* To answer this question, one would need to access data deemed to be the "current" era of NBA playstyle. This has been regarded as having begun around 2009/2010, so to be safe, the decision to test data on the last ten available years of NBA seasons was decided upon. The data for this project was collected as a .CSV file that was found with the help of a QAC tutor during office hours. The website it was found on was Kaggle, in which the author, user Michael H., scraped data on each NBA season from every team between 2000-2022 on stats.nba.com for regular season and playoff stats. The intention is to trim that dataset using appropriate management techniques to range from the years 2010-2021.

In total, there are 329 Observations over 11 seasons and 30 teams, meaning there is one missing observation in the entire dataset. Variables worth noting in this dataset are all numeric and specific variables including FG. (Field Goal %), FT. (Free Throw %), REB. (Def Rebound + Off. Rebound), TOV (Turnovers). These four variables are the independent variables from which Oliver's 4 Factors emerge. Additional statistics that appear to be significant in the study include 3P., which is just the 3-point shooting percentage of a team, and STL, which is the average number of times a team steals the ball from the other throughout the season. The dependent variable worth noting is WIN., which is just Wins/Wins+Losses.

The procedure in which the data will be analyzed will be from first applying the correlation tests to each variable of consideration in this study on variables to get a better understanding of their importance to an NBA team's performance throughout the regular season. Basic plots using ggplot will then be run on each of the Independent variables in conjunction with the dependent for analysis. From there, the obvious method of analysis present in this study is the linear regressions that will be run on the Dependent Value (Wins/Losses) and the 4 Independent Values, from which, they are to be compared to new combinations of variables to then interpret which of the models incorporating newly emerging statistics like 3P% or STL has the highest adjusted R^2 and lowest AIC value; all this to conclude whether Dean Oliver's breakdown of the importance of winning (FG%, TOV, Rebounding, and FT%) still stands. Next, supplemental statistical analysis methods include all four cross-validation methods: LOOV, K fold, K fold repeated, and Holdout, and from there, deriving which provides the lowest RMSE and highest adjusted R^2. Clustering will help to provide insight into which observations are similar to one another, and after creating a cluster, studying the different averages of studied variables in each cluster to derive some further insight into each cluster and their characteristics worth noting. The last form of supplemental statistical analysis run is PCA, which will give insight to which variables of measurement provide similar insight to one another. This step can be key to understanding which variables to include in regression analysis and which to exclude.

Some limitations and points of consideration are the emergences of new advanced statistics or new findings in the past 20 years that have put Oliver's Four Factors in a new light. An example of this comes from John Ezekowitz of the Harvard Sports Analysis Collective, where, in 2011, he discovered that the statistic of Free Throw Made/Possession was as good, if not, better of a variable to measure wins over FT%, with the T-test application and regression leading to better results.

Results: Of the basic correlation tests run as a preliminary to the regressions, it was apparent that the inclusion of the 3P% was the right idea, as it had an extremely high correlation value with WIN%, second only to FG%. Another observation worth noting is that OREB was negatively correlated with WIN%, an important observation worth noting. The process of creating models of multiple linear regressions was one of utilizing intuition to reach a sound conclusion. First, adding the 3P% to the 4 factors and then replacing each factor with 3P% to determine which variable was least supportive to WIN% out of the 4 factors, the answer being FT%. 3P% emerged early on as an extremely important variable in predicting WIN%. Additionally, the surprising appearance of STL as an important predictor of WIN% emerged in the regression modeling. This is stated because the multiple models ran incorporating an assortment of variables, the one with the lowest AIC including both 3P% and STL. Additionally, another important finding of the regressions was that FT% was not an important predictor of WIN% and rather led to a higher AIC and lower adjusted R^2 value than without it. After that, the next step was to check for moderator presence by creating new variables that were the product of two other variables. There was no presence of moderators detected in the given dataset. After plotting the best model the residuals did not have a mean of 0, unfortunately, meaning heteroskedasticity.

After clustering, 4 clusters were determined after using the elbow method. After that, observations were made on each cluster and their statistical drivers. All the clusters except cluster 3 were similar to one another in terms of WIN%. Cluster 3 had a far lower 3P% than either of the 3 other clusters. That being said, this trace could be found in the other statistical averages of different variables of each cluster. Cluster 3 had the lowest 3P%, FG%, and FT%, as predicted after analyzing WIN% of each cluster. The concentration of cluster 4 appeared to be the greatest, followed by 3, 2, then 1. Additionally, most of the teams from the earliest seasons of data included are clustered in 2 and 4, the midpoint, around 2016/17, had a wider distribution of clusters, but mainly cluster 4. The latter half of the data, more recent seasons, are clustered under 3 and the presence of cluster 1 is scattered throughout the data and isn't quite concentrated at a given time.

PCA provided the insight that the variables FTM, FTA, and PFD are very similar, an unsurprising finding since all relate to free throws. Additionally 3PA, 3PM, DREB, are FGM very similar, likely because of how prominent of predictors each is to WIN%. BLK and STL are also very similar, which is interesting and not surprising since both are prominent defensive statistics. Overall, 7 deduced principle components were present on the variable list given.

All four cross-validation methods were run and the one that came back with the lowest RMSE and highest adjusted R^2 was the K-fold method, with values of 0.06133163 and 0.8158944 respectively. Then followed by Holdout, then far behind that is LOOV, with a big gap between 2nd and 3rd, and finally, K fold repeated with the worst returns.

Discussion: It is safe to say that Dean Oliver's regressions can be stated as outdated to today's current NBA playstyle. The inclusion of the 3P% statistic alone was far greater of a weight in predicting WIN% than FT% for example, the weakest of the originally posed Four Factors. To answer the first of the two research questions: *How important are Dean Oliver's Factors to winningness?* The answer lies in the first of the multiple linear regression that was run where it was found that the original four factors gave an adjusted R^2 value of .4927 and AIC of -525.39. Other regressions explained more of the variation of WIN% and had a lower AIC, the model selected, for example, had an adjusted R^2 of 0.5822 and AIC of -588.26. The second part of the research question, where it is asked whether there are variables not included in the original regression that should be has also been answered by the selection of the best model available, which includes the 3P% and STL variables. These two variables are unmentioned in Oliver's publishings and replace the FT% statistic to get the highest achievable adjusted R^2 and AIC values.

It's important to discuss what more should be done to produce a sound publication worthy of recognition in the sports statistics world. First, the observations are lagging by around two years, so to fully capture the current period, updated observations including the last two seasons would be incredibly helpful. Secondly, the inclusion of more advanced basketball statistics that are more recent will likely have a strong relationship with WIN%, due to being a product of the time. So for more advanced research, I would consider the usage of all sorts of statistics to see if any other sneaky predictors help predict WIN%, but to do so, would have to navigate for scraped data online. Additionally, to pose any validity to my findings, I would need to ensure my regression analysis had a residual mean of 0, which would be a future step were I to continue this project. Another potential step would be to include the caret package and perform different forms of regression analysis.

There are limitations of this study that if addressed, would make for a better and more sound publication. The first is that pegging WIN% to just variables in this study and not the multitude of other factors that go into a successful NBA game is tough but what Data Analysis has to deal with, the intangibles are there and are hard to capture in this case. Another impact of COVID lies in the form of interrupting the hegemony of the intangibles that lie in the game of basketball, which is something worthy of consideration. The NBA bubble was an experiment in which all 450 NBA players would be in a bubble environment and play basketball in closed stadiums with no in-person fans and no home-court advantage. This switch-up of the environment has called into question the NBA champions of that season, the Los Angeles Lakers, and whether that championship counts or not. Although, very recently all four teams that reached the conference finals during the NBA bubble recently reached the conference finals once again in the 2023 season, which, to many, puts to bed the rhetoric that the Bubble was easier or a fluke.

Overall, one could take this research project and decipher that, although there is much more work to be done before publishment, there are some key takeaways that call into question aged statistical analysis on the sport of basketball and implore statisticians to use modern data to research the nature of the sport of basketball, due to its extremely fluid nature of playstyle. Dean Oliver's Four Factors, although outdated, presented some key observations at the time of its publication on the statistics that drive winning NBA basketball. That is not to say, however, it does so successfully in today's game.

Bibliography:

- Chatterjee, S., Campbell, M. R., & Wiseman, F. (1994). *Take that jam! an analysis of winning percentage for NBA teams*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/mde.4090150514
- Jacobs, Justin. "Introduction to Oliver's Four Factors." *Squared Statistics: Understanding Basketball Analytics*, squared2020, 5 September 2017, https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/. Accessed 11 April 2023.
- Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019, October 7). *Trends in NBA and Euroleague Basketball: Analysis and comparison of statistical data from 2000 to 2017*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0223524
- Zając, T., Mikołajec, K., Chmura, P., Konefał, M., Krzysztofik, M., & Makar, P. (2023). Long-Term Trends in Shooting Performance in the NBA: An Analysis of Two- and Three-Point Shooting across 40 Consecutive Seasons. *International journal of environmental research and public health*, *20*(3), 1924. https://doi.org/10.3390/ijerph20031924