What Goes into Winning an NBA Game?

A Statistical Research Project by Harrison, Charlie, Alan, and Alex

Project Introduction

Winning is King. Our group, full of enthusiastic sports fans, wants to understand which predictors are the most influential toward winning. We hope to quantify what exactly “influential” means (particularly, in comparison to other possible predictors).

Going into this project, we believe that true shooting percentage, blocks per game, points-per-game, three-point shooting percentage, and turnover rate are important to winning, but we want to gain some quantitative insight into how influential these factors actually are towards winning (or perhaps if they are influential at all). Similarly, we can analyze how these factors might correspond with winning championships and making the playoffs. We will also try to see if other predictors (those that we might not have realized) are indeed strong indicators of a winning team. We also can see if joint predictors (i.e. blocks per game + true shooting percentage) are substantially more predictive than their components. Note: we will be looking solely at team statistics, not player statistics.

Our plan is to use multiple linear regression models to understand which predictors have the biggest impact on winning. Initially, we will be using excel to help sort and organize datasets from the last 5 seasons, so that we can isolate the fast-paced, offensive style of game in the NBA that has become a recent phenomenon. We will then convert the data from these excel sheets into Pandas data frames for use in Python. Next, we will analyze the coefficients of the linear regression analysis, while taking into account correlation/association among the predictors. We will also look at the p-values to make sure our coefficients are statistically significant. We do many different linear regressions in order to be able to isolate certain predictors and analyze other similar questions like how a team's previous performance affect their future one. We will also explore other possible predictors of the winning rate of a team for a given season.

Defining Key Terms

- Acronyms used:
    - PTS (Points)
    - PM (Shots made)
    - PA (Shots attempted)
    - G (Game)
- Field goal attempts (FGA): Number of shots attempted by a team.
- Free throw attempts (FTA): Number of free throws attempted by a team.

- True shooting percentage (TS%): A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.[1]
    - TS% = PTS / (2 * (FGA + 0.44 * FTA))
- Points per game (PPG): The average number of points that a team scores in a game (averaged over a season of games).
    - PPG = PTS / G
- Blocks per game (BPG): The average number of blocks that a team has in a game (averaged over a season of games).
    - BPG = BLK / G
- Percentage of three-pointers shot (3PA): The percentage of a team's field goal attempts that are three-pointers.
    - 3PA = 3PA / FGA
- Three-point shooting percentage (3P%): The percentage of a team's three-pointers attempted that it makes.
    - 3P% = 3PM / 3PA
- Margin of victory (MOV): The average margin of victory (or loss) for a team over the course of a season.
    - MOV = (Points For) - (Points Against)
- Turnovers (TOV): Number of turnovers per game over the course of a season.
    - TOV = Seasonal turnovers / G
- Assists per game (APG): The number of assists a team gets per game over the course of a season.
    - APG = # of assists / G
- Two-point shooting percentage (2P%): The percentage of a team's 2-point field goals that are successfully attempted.
    - 2P% = 2PA / FGA
- Free throw percentage (FT%): The percentage of a team's free throws that are successfully attempted.
    - FT% = FTM / FTA

Methods

The goal of our project is to find how particular statistical metrics contribute to team success in the NBA. To answer this question, it is important to get the newest dataset possible so our results are reflective of the current game. We chose a five-year time period from the 2017-2018 season to the 2021-2022 season. We thought five years was a sufficient sample to find meaningful correlations. To find the key data items identified above, we scraped Basketball Reference's seasonal team data sets and uploaded the resulting seasonal data to GitHub. We also constructed our own categorical outcome variable (playoffs) which has three separate result

[1] https://www.basketball-reference.com/about/glossary.html#tsa

options depending on how each term performed: playoff berth, finals berth, or no playoff appearance at all.

Once we collected our data in GitHub, we began our code by grabbing the data and converting the important objects into floating or integer numbers depending on what we'll need to analyze.[2] Then, we calculated the values of the certain key terms that we defined above.[3] Next, we sorted all of the teams in our dataset by winning percentage to get an idea of the winning percentage range.[4]

From there, we examined teams by year (for example, we separated the 2021 Warriors data from the 2022 Warriors and so on). We expected true shooting percentage, blocks per game, and points per game to be our most important predictors based on our naive personal eye test of basketball. Our next step was to construct a series of linear regression models using some of our various statistical metrics. The first three models use a combination of various statistical items as metrics to predict winning percentages in the regular season. For each regression, we first created a residual plot and then printed the summary of the results. We had a particular interest in the coefficients as they told us which predictors are most influential and R^2 which told us how well our model succeeded at predicting team wins. After running these quantitative regressions, we used a model to look at the relationship between the categorical variables of playoffs defined above and the MOV of the team. We were later curious about the relationship between the previous year's results versus the current year's. So, we also constructed an autoregressive model which asked how well wins in years prior were able to predict the results of wins in the following year.

Based on the regression models performed above, we identified true shooting percentage (TS%) to be a particularly powerful metric. We wanted to know if we could further enhance its predictive power by adding eight additional features. We took TS% and put it to the power of 2 and made that a feature and we repeated this process all of the way up to 9 and made each of those results a feature. Then, we printed our results and compared the R^2 to the naive model based on just TS%.

As a curiosity, we ran one final regression which considered how the exclusion of a constant impacted true shooting percentage's significance toward winning.

All of our code can be read and run in this colab notebook.

[2] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=vDbAj9WADdz5&line=2&uniqifier=1

[3] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=bUFB6042Dseq&line=1&uniqifier=1

[4] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=GaBomgbt4xts&line=1&uniqifier=1

Results

The four multiple-variable linear regressions revealed the discrepancy between statistical metrics in predicting basketball wins. In NBA sports media, there is a strong emphasis on the so-called big three metrics: points per game (PPG), blocks per game (BPG), and assists per game (APG). These are volume metrics because they just relate to the number of items achieved, not the efficiency. Our analysis found that this emphasis may be overstated and there should be a greater focus on efficiency metrics. Efficiency metrics indicate how accurately a player scores a ball. We drew this conclusion from two separate regressions, where one used the big three as features and the other focused on efficiency metrics (2-point shooting percentage, 3-point shooting percentage, and free throw percentage). In the volume-metric stats regression, we found 0.0168 is the expected increase in winning percentage if PPG goes up by one assuming all other predictors are fixed.[5] This number is lower than the coefficient for BPG which has an expected difference of 0.0209 with all predictors fixed. This does not mean that BPG is more important than PPG. We did not standardize these metrics, so since the number of blocks in a game varies much less than the number of points in a game, we cannot assume which one is more significant. To do so, we can multiply the coefficients by their statistics standard deviations, and then compare them on a fair scale. With this, we see that a change in one standard deviation in PPG leads to a 0.073 increase in win percentage, versus a 0.015 increase from an increase of one standard deviation in BPG. Additionally the p-value for blocks per game – 0.171 – is higher than our significance level of 0.05. This means we can not decisively reject the null hypothesis that BPG has no impact on winning. In future analysis, we would like to look at a larger sample of years to clarify the impact of BPG. Interestingly, APG has a basically zero coefficient as a predictor and contributes nothing to R^2. This takeaway is supported by the high p-value of 0.983. The significance of this result is further discussed below.

The efficiency-metric based linear regression had a much higher R^2 than the volume-metric based one: 0.52 vs 0.3.[6] The most influential predictor in the efficiency-metric regression was three-point percentage. This makes sense given the playstyle of the modern NBA. Teams are more focused on three-point shots, and they now more than ever dramatically impact the outcome of a game. Overall, the predictiveness of efficiency metrics demonstrates a misalignment in the fan/media perspective of the game and the statistical story. To further support this thesis, we found the regression that just focused true shooting percentage (TS%) had a higher R^2 than the one using three of our volume metrics (0.492 vs 0.3)[7]. The fact that R^2 for TS% is lower than the three-feature regression referenced above might imply that the formula for TS% is suboptimal and could be tweaked to improve its predictive power. We tried our hand at modifying our true shooting percentage model by creating eight other features based on TS%. Adding the features did modestly up the R^2. Ultimately, TS% is a static formula and because

[5] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=jTX57aqiD2Cr&line=1&uniqifier=1

[6] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=tfdKX43nEU66&line=1&uniqifier=1

[7] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=hAa2_iXwEnoh&line=1&uniqifier=1

teams have a variety of shot diets, it's not surprising that it does not capture the whole story of team efficiency.

Another interesting finding from our quantitative regressions is the relative lack of predictiveness from playmaking-related stats.[8] Playmaking in basketball is the act of helping another player score without directly doing the scoring yourself. The regressions for turnovers per game and assists per game each had below 0.1 R^2 individually and even when used as features together with another stat, three points attempted percentage, they still had a fairly low R^2 of 0.2.[9] This does not mean these stats are useless, but they should be paired with a variety of other metrics when evaluating players. Clearly, they have little predictive value when being used individually to assess players. The p-values for the coefficients of TOVpG and ASTpG were both rounded to 0 by Python though, so they likely have some effect on how winning a team is.

We also were curious about how making it to the playoffs and having success in the playoffs relate to the margin of victory (MOV). Unsurprisingly, playoff success is correlated with MOV, with an R^2 of 0.649.[10] We would think that teams that perform well in the playoffs would score more than the points they give up. That said, the imperfect association is likely attributable to teams that often blow out their opposition, but are not always the most consistent. The 2020 Mavericks had this same issue where their offense was incredibly explosive, but they suffered from depth issues that were exposed in the playoffs.

All of the aforementioned regressions we ran attempted to predict winning percentage, except for the categorical regression. Based on the residual plots, these regressions generally abided by the linear regression model assumptions of normality; however, the categorical regression could be interpreted as not satisfying these assumptions. When the plus-minus (MOV) values were extreme, most of the values were above the reference line or most were below the reference line. Other than this linear regression's residuals, there was mostly an even number of positives and negatives, implying the expected normality of errors. The plots often had a sort of parallelogram shape though, signifying a lack of randomness. This implies that not all of the assumptions of linear regression are captured in our models, but we will still consider what the models can inform us about winning predictors.

Another intriguing question is about how previous success predicts future success in the NBA. To answer this question, we created an autoregressive model. Unsurprisingly, previous performance did positively predict future performance with an R^2 of 0.267.[11] Teams usually keep most of their players year after year, so this tells an obvious fact that results are not random in the NBA. This is likely attributable to good drafting, strong player improvement, better coaching, etc. These

[8] Note: we already saw this above with APG's lack of predictive power vs PPG and BPG

[9] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=hmKPxLRTEBoN&line=1&uniqifier=1

[10] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=xn3avEAx9H8N&line=5&uniqifier=1

[11] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=98QEmKhdEiX0&line=3&uniqifier=1

variables are not constant, which is why results vary from year to year. Confirming the reliability of this autoregressive model, the residual plot strongly demonstrates the normality of errors in the predictiveness of the model. It also is randomly distributed indicating that the assumptions of linear regression are satisfied.

As an interesting aside, we ran a regression on true shooting percentage and winning percentage with and without a constant. This led to us finding an interesting statistical fact not covered in class. We saw from the residual plot that it failed the normality criteria of linear regression, but it gave a very high R^2.[12] We have learned that this R^2 increase is a well-known effect in statistics, called an uncentered R^2 since it does not include a centering intercept. As we expand our knowledge of the field of statistics and data science, we hope to understand this idea even better.

Discussion

In the end, we were able to answer some of the questions that we set out to answer. We found that within a team, the best metrics to determine the chances of winning are the scoring efficiency metrics. We didn’t find the holy grail of all metrics as we looked at things like true shooting percentage, blocks per game, points-per-game, three-point shooting percentage, and turnover rate. However, we did find that when mixing a few of these metrics, more specifically the model using the two-point, three-point, and free-throw shooting percentage together and the model using the true shooting percentage, we found the highest r-squared values.

In the future, we could compare an analysis of the “bubble year” – the 2019-2020 season where the NBA played in a “bubble” due to Covid-19 – with the results from other, “normal”, seasons. The reason this could be compelling is that the stats from that year were different due to play being in a crowdless setting. Furthermore, we could conduct more train and test analysis (where we train from statistics from certain seasons and then test on other seasons), which we didn’t do in this project. We could also use more advanced statistics in the future such as advanced plus-minus statistics that dictate how impactful a player is to their team while they are on the court. By considering additional advanced statistics, and using a combination of factors that is beyond our scope of analysis, we could find potentially more predictive statistics than true shooting percentage. A final direction we could go in would be to construct prediction metrics based on player statistics, instead of just team statistics. For example, we could fill the shoes of a modern-day GM and try to model how different combinations of skills of players could affect the competitiveness of teams. This would be an incredibly complicated project and is likely what many professional basketball data analysts do on a daily basis.

[12] https://colab.research.google.com/drive/1I0wVo6k3vElRpioCs13FHwGZ6uv7WvXM#scrollTo=tmAIAjAbEv6O&line=1&uniqifier=1

Bibliography

“2021-22 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2022.html#totals-team.

“2021-22 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2022.html.

“2020-21 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2021.html#totals-team.

“2020-21 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2021.html.

“2019-20 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2020.html#totals-team.

“2019-20 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2020.html.

“2018-19 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2019.html#totals-team.

“2018-19 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2019.html.

“2017-18 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2018.html#totals-team.

“2017-18 NBA Season Summary.” *Basketball Reference*, 2022, https://www.basketball-reference.com/leagues/NBA_2018.html.