Money in the NBA: Analyzing the Factors to Win

Alexander Martinez

University of Arizona - Tucson

Abstract

Using a dataset found on a web-based data-science environment, *Kaggle,* I created two different models that describe the win probability of NBA teams from the 2003 – 2021 seasons. First, I will hypothesize the important implications this has for the team managers, the gambling industry, and the National Basketball Association alike. I will then break down these models by first highlighting the assumptions that were made while analyzing these models. Once the assumptions are described, I will then describe the first regression model I ran on the data found for the teams that played at home from the 2003 – 2021 seasons, and, after, I will do the same for the away teams. This paper will then be concluded by key take-aways that managers of the league and the gambling industry may find important.

Point of Research – An Introduction

The National Basketball Association was first founded in 1946, and it has been a great success since then. More companies, such as sponsors and gambling companies, are becoming more invested in the league as it continues to explode. Since 2001, the league's revenue is up 297%. In the 2020 season, all 30 organizations in the league generated a combined revenue of 7.92 billion US dollars. (Gough, 2021) For this reason, there has been an influx of sponsorship partners eager to take a slice of the cake. Since the 2010 season, each season has made more sponsorship revenue than the previous year. As of 2021, sponsorship revenue in the NBA is up 272% from 2010, a 900,000,000 million US dollar increase in the last decade. (Gough, 2021) The implications of phenomenon can be split into two equally important perspectives: gambling and managerial.

To start, with the sports gambling industry booming, we can expect to begin to see an even greater viewer investment across the United States. On any given night, commercials are starting to advertise companies like *FanDuel* or *Draftkings*. This aligns with a trend that the United States has been experiencing recently: 41% of sports bettors started sports betting in the last year alone. (The Harris Poll, 2021). Sports wagering companies are making profits from a booming demographic, which directly presents the National Basketball Association with an opportunity to attract more viewers. The models that I will outline further into the paper does have uses for gambling, such as predicting a team's win probability by inputting specific game statistics that are relevant to my model. This can be applied if your team is home (Model 1) and/or away (Model 2).

From an organization's managing perspective, one notices that the NBA is on a clear path to maximize its revenue, but managers are more concerned on the who rather than the what. They want to know who will benefit from the potential revenue the NBA receives. Small Market teams such as the New Orleans Pelicans, who raised 209 million US dollars in revenue will presumably not acquire as much capital as the Golden State Warriors who generated a grand total of 474 million US dollars in revenue. This is an extremely important connection to make, because, in the last decade, 9 out of the 10 teams who generated the most revenue in the NBA have been a part of or won a league championship and/or conference championship. (Gough, 2021) In other words, teams that win have a strong connection to receiving and generating revenue. There needs to be a way to predict win probability and observe specific in-game statistics that play a significant role in winning, which will, in turn, benefit the team's financial success as well.

In this paper, I will examine a model that will predict the win probability for away teams and for home teams, separately. They will also be used to describe which relationships seem to be significant for a team to win.

Relevant Work

Though there is a lack of research done that focuses mainly on the connections between revenue and winning, I did find that there was an exceptional publication ahead of me that explored similar routes. In particular, "Application of machine learning on NBA data sets" by Jingru Wang and Qishi Fan. In their paper, they use back-test using decision tree, random forest, and gradient boosting models, and they use them to try to make all-star predictions, and test if *Hot Streak* is, in fact, a fallacy. Most notably, they do try to make playoff predictions, and they found, according to their data, that field goal percentage, that is, the percentage of attempts that are converted into a scoring point, appears to be the most important team statistic when it comes to making the playoff picture. As a reference, to make the playoffs, you must win more than enough to be in 10^{th} place or higher in your respective conference – East or West, of which each are made of 15 teams. The idea that they found this statistic to have significant effect on something that is only possible with a team win (making the playoffs) will show to be consistent with a portion of what my models exhibit.

Dataset Retrieval and Description

The dataset was retrieved from Kaggle, a well-known data science website filled with a community of like-minded, statistics-oriented interests. The software I had used was *R* on the source-code editor, *R Studio,* and the data was originally recorded directly by the NBA. I had chosen this dataset because the sample size was large and it contained several statistics that were

already solely team statistics, which meant they were consistent with what they represented. Since the original data recorded every single game from 2003- 2021, the dataset originally consisted of ~44280 observations (82 games × 18 seasons × 30 teams). The original dataset had consisted of many variables for each game starting at the beginning of the 2003 season. The data was then condensed down into two master datasets (Home and Away) with ~540 observations (30 teams × 18 seasons) each, when I took the mean of each team statistic and then divided them out by team and season. For example, the Atlanta Hawks have 1 observation from the home-side of the 2003 season that has their average team statistics all for that same season, and they have one for the away-side of the 2003 season. This can be visualized below:

	SEASON	TEAM_ID_home	avg_points	win_prob	avg_fgpct_home
1	2003	ATL	95.78571	0.4090909	0.4385000

Figure 1: Atlanta Hawks 2003 season averages when they were at home

	SEASON	TEAM_ID_away	avga_points	win_probaway	avg_fgpct_away
1	2003	ATL	89.46341	0.3260870	0.4268537

Figure 2: Atlanta Hawks 2003 season averages when they were away

Predictor and Response Variables

To start, the dataset did not contain variables that are near impossible to record, such as team chemistry, player fatigue, player attitude. It also did not contain variables that I personally would find to be significant, such as the miles traveled for an away game the night before, whether they were facing a team without their star player in the regular rotation of the game. Nonetheless, contained in the data were undeniably necessary predictor variables for each team

and their respective season such as: average points per game, average field goal percentage per game, average 3-point percentage per game, average assists per game, average free throw percentage per game, and finally, average rebounds per game. The response variable, of course, was the win probability of a team, which was set as a percentage. Also, as a final note, there were some observations that were completely removed due to some irregularities in the dataset, so it is not a perfect 540 observations, as some were nulled.

Assumptions

Assumption 1 is that each of the observations are independent of one another other. Now, this assumption fails because there are error terms that may be correlated with x's like mentioned before, such as player fatigue, coach likeability, etc. In order to fix this, we would need and randomized experiment. Assumption 2 is that all variables are identically and independently distributed. Assumption 2 may possibly fail because there seems to be correlation between average field goal percentage and average free throw percentage. The correlation level is at .4056. Assumption 3 states that there is variation $X > 0$. The data has about 530 observations, which is much more than the number of variables, so this assumption passes. The last assumption is that a 4^{th} moment exists in my observations. This assumption passes because none of the outliers are approaching infinite.

Models

The formula for the model is shown below. As we can see, each part is broken down into matrix form in order to maximize understanding of the model that I have used.

$$Y = X\beta + \epsilon$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Figure 3: A Multiple Linear Regression Model Shown in Matrix Form.

Credit for Photo: https://www.ctspedia.org/do/view/CTSpedia/LinearRegression

As mentioned, my first model, Model 1, tries to predict the win probability for the home team. The predictors, as can be seen below, were all included, besides average assists per game. Average assists were taken down because it seemed to lack significance at the $\alpha = .01$ level to be included in this model. The summary of the model, without average assists, in order, from top to bottom, includes average field goal percentage per game, average free throw percentage per game, average three-point percentage per game, average rebounds per game, and average points per game. Again, these are not the averages of their total; these are the averages of their home games. The model summary is shown below

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.30422 -0.07183  0.00120  0.07376  0.34374

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.612725   0.228764 -15.792  < 2e-16 ***
X1_avgfgpct   6.609224   0.418272  15.801  < 2e-16 ***
X2_avgftpct   0.566366   0.172562   3.282   0.0011 **
X3_avgfg3pct  1.426163   0.260453   5.476 6.72e-08 ***
X4_avgreb     0.031281   0.002943  10.629  < 2e-16 ***
X5_avgpts    -0.011061   0.001219  -9.072  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1151 on 533 degrees of freedom
Multiple R-squared:  0.4772,    Adjusted R-squared:  0.4723
F-statistic:  97.3 on 5 and 533 DF,  p-value: < 2.2e-16
```

Figure 4: Home Regression Summary

The response variable was win probability shown as a percent. The adjusted R-squared is .4723, which indicates about 47.23% of the variation in the response variable in the model is explained by the variation in the x variables. I also tested the model and found that home teams, given the averages of all their season home stats combined over the 2003 -2021 seasons, have a win probability of 57.41%. The model is useful to predict a team’s home court winning percentage in any given season.

For the second model, Model 2, I had the same exact predictor variables, but, obviously, this was the away winning percentage regressed against the away season average team stats from each team observed.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.28409 -0.07844 -0.00005  0.07932  0.32891

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.274383   0.225349 -14.530  < 2e-16 ***
X1A_avgfgpct   5.201412   0.452914  11.484  < 2e-16 ***
X2A_avgftpct   0.584163   0.165581   3.528 0.000455 ***
X3A_avgfg3pct  1.559034   0.257521   6.054 2.67e-09 ***
X4A_avgreb     0.026675   0.002951   9.039  < 2e-16 ***
X5A_avgpts    -0.007617   0.001214  -6.277 7.17e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 533 degrees of freedom
Multiple R-squared:  0.3929,    Adjusted R-squared:  0.3872
F-statistic: 68.98 on 5 and 533 DF,  p-value: < 2.2e-16
```

Figure 5:Away Regression Summary

Notice the similarities between Model 1 and Model 2. The same respective variables were significant. This indicates that these statistics are in fact big enough factors when it comes to winning, no matter if the team is home or away. Also, the adjusted R – squared on the away side indicates that only about 39.29% of the variation in the response variable is explained by the variation in x. I, again, tested the model and took the away averages across the ~540 or so observations and inputted these to the model. The result was that if a team were to get those exact averages in a whole season, while being away, they would have a win probability about 40.33%.

These probabilities are consistent with articles, such as “How important is home-court advantage in the NBA?” by (Kelvin Blehumeur, 2017) and many others that state the home team wins about 60% and the away team wins about 40% of the time. This further reinforces the model’s ability to predict with accuracy.

Conclusion

All in all, I am satisfied with both my models and their abilities to predict, with accuracy, the win probability of the Away Teams and the Home teams. This does lead to the part of how these models can be useful for real-world, practical use. Although these models do not pass at least 1 important assumption, they do seem to be reinforced by outside research and investigation. I will separate the key takeaways into two perspectives, just like I had done at the beginning of this paper: gambling and managerial.

From a managerial standpoint, it is important to restate that history has shown (refer to paragraph 3 in Introduction) that teams who have championship appearances or wins, tend to create more revenue. The main reason I had separated the regressions into home and away was to see if there were any difference in which team's statistics were significant, and I wanted to be able to highlight what matters at home and away. Surprsingly, they came back with similar results. As we saw from the models, the same statistics seem to be significant when it comes to effecting win probability. As advice to those team managers, in order to win games, they should shift their primary focus to practicing on maximizing their value on those specific statistics, such as average field goal percentage, average points etc. If they can do that, the models predict that their win probability will go up. This then translates to having a higher chance at making a substantial playoff run. As a result, more fans become interested, and more revenue is generated.

From a gambling perspective, the models can be used with rough accuracy, to predict the amount of wins a team may have by the end of the season. A gambler may find it intriguing that the models like mine can be useful to make their probability of winning their bets increase slightly. Gambling companies, on the other hand, can start creating models like the ones I have created to use as a selling point to get more people to gamble on NBA games. This will not only

create more revenue for betting websites also for the NBA, as gamblers increase their interest in watching the NBA.

All in all, the NBA is already exploding and growing rapidly. They might find that there are many correlations to winning and revenue, which can help them create more money.

Citations

Gough, C. (2021, February 17). *Total NBA revenue 2001-2018*. Statista. Retrieved December 8, 2021, from https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/.

Gough, C. (2021, September 28). NBA sponsorship revenue 2010-2021. Statista. Retrieved December 8, 2021, from https://www.statista.com/statistics/380270/nba-sponsorship-revenue/.

Gough, C. (2021, February 17). NBA Teams Revenue Ranking 2020. Statista. Retrieved December 8, 2021, from https://www.statista.com/statistics/193704/revenue-of-national-basketball-association-teams-in-2010/.

Feider, M. (2021, January 13). Have sports viewership and sports betting been impacted by the pandemic? The Harris Poll. Retrieved December 8, 2021, from https://theharrispoll.com/have-sports-viewership-and-sports-betting-been-impacted-by-the-pandemic/#:~:text=However%2C%20betting%20on%20sports%20increases%20viewership%20and%20enjoyment%20among%20sports%20bettors.&text=63%25%20of%20these%20sports%20bettors,have%20bet%20on%20the%20outcome.

Wang, J., & Fan, Q. (2021). Application of machine learning on NBA data sets. Journal of Physics: Conference Series, 1802(3), 032036. https://doi.org/10.1088/1742-6596/1802/3/032036

Belhumeur, K. (2017, October 3). *How important is home-court advantage in the NBA?* Bleacher Report. Retrieved December 8, 2021, from https://bleacherreport.com/articles/1520496-how-important-is-home-court-advantage-in-the-nba.