

NBA HOME WINS PREDICTION

Langxi Wang

The Bolles School, Jacksonville, USA

DOI: 10.46609/IJSSER.2022.v07i02.014 URL: <https://doi.org/10.46609/IJSSER.2022.v07i02.014>

Received: 7 Feb. 2022 / Accepted: 19 Feb. 2022 / Published: 28 Feb. 2022

ABSTRACT

Aim: This study aimed to build a predictive model for NBA Home Wins using logistic regression.

Method: A public data was used in this study. All the records were randomly assigned into 2 groups: training sample (50%) and testing sample (50%). Logistic regression was built using training sample: artificial neural network and linear regression.

Results: Home teams won 1474 of 2230 games, or 66.1%. According to the logistic regression, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away, sum_PF_home, sum_PF_guest were significant predictors of the winning of the home teams. The area under curve was 0.9815. The optional cutoff time is 0.52. The mis-classification error was 0.06. The sensitivity rate is about 95.4% and the specificity is 90.4%.

Conclusions: In this study, we identified important of predictors of NBA Home Wins in the United States, for example, % of made shots and total of personal fouls. This tool will be very helpful to understand features determining NBA Home Wins and to maximize the likelihood to win.

Literature review: Before researching my topic, I also referred to many other similar experiments. These included Fadi Thabtah, Li Zhang & Neda Abdelhamid on the use of machine and artificial intelligence to learn and analyze historical data to predict the likelihood of a team's victory. To achieve these goals, the article selects several machine learning methods that use different learning schemes to derive models, including Naive Bayes, artificial neural networks, and decision trees, among others. Several learning schemes were used to finally arrive at which NBA team would win.

1: Introduction

From the expansion of the postseason bracket to 16 teams in 1983-84 through the 2018-19 season, home teams won 61 percent of their games in the regular season and 65 percent of their games in the playoffs. That pattern holds even accounting for the fact that better teams play more home games in the playoffs.ⁱ

In team sports, the term home advantage – also called home ground, home field, home-field advantage, home court, home-court advantage, defender's advantage or home-ice advantage – describes the benefit that the home team is said to gain over the visiting team. This benefit has been attributed to psychological effects supporting fans have on the competitors or referees; to psychological or physiological advantages of playing near home in familiar situations; to the disadvantages away teams suffer from changing time zones or climates, or from the rigors of travel; and in some sports, to specific rules that favor the home team directly or indirectly. In baseball, in particular, the difference may also be the result of the home team having been assembled to take advantage of the idiosyncrasies of the home ballpark, such as the distances to the outfield walls; most other sports are played in standardized venues.ⁱⁱ

The term is also widely used in "best-of" playoff formats (e.g., best-of-seven) as being given to the team that is scheduled to play one more game at home than their opponent if all necessary games are played.ⁱⁱⁱ

This study aimed to build a predictive model for NBA Home Wins using logistic regression. Most of the individual data for this study, as well as team data, came from <https://www.basketball-reference.com/>. This contains detailed and comprehensive comparisons between the various stats through various categories and percentages. Among other things, I have summarized some of the factors that affect a team's winning percentage by comparing the team's winning percentage to their other individual stats and analyzing them.^{iv}

2 Data and Methods:

Data

This dataset is public available for research. It is available here:

<https://www.kaggle.com/nathanlauga/nba-games>.

The outcome of interest is called HOME_TEAM_WINS (1=Yes; 0=No).

FG_PCT_home/FG_PCT_away: % of made shots (excluding free throws)

FT_PCT_home/FT_PCT_away: % of made free throws

FG3_PCT_home/FG3_PCT_away: % of made 3-point shots

AST_home/AST_away: # of assists

REB_home/REB_away: # of rebounds

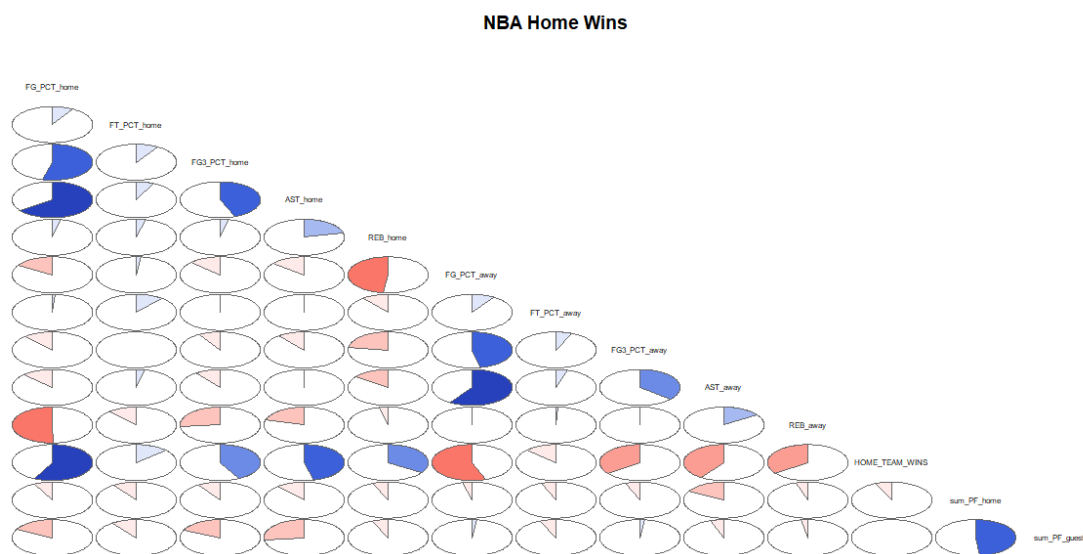
PF_home/ PF__guest: personal fouls

All the records were randomly assigned into 2 groups: training sample (50%) and testing sample (50%). A logistic regression model was built using training sample.

3 Results:

Home teams won 1474 of 2230 games, or 66.1%.

Figure 1: matrix of correlations between variables



This diagram shows the relationship between each of the variables. Where blue indicates positive correlation and red indicates negative correlation. The darker the blue color, the closer the positive correlation between the two variables is to 1, that is, the greater the positive correlation, and vice versa, the smaller the positive correlation. Similarly, the darker the red color is, the

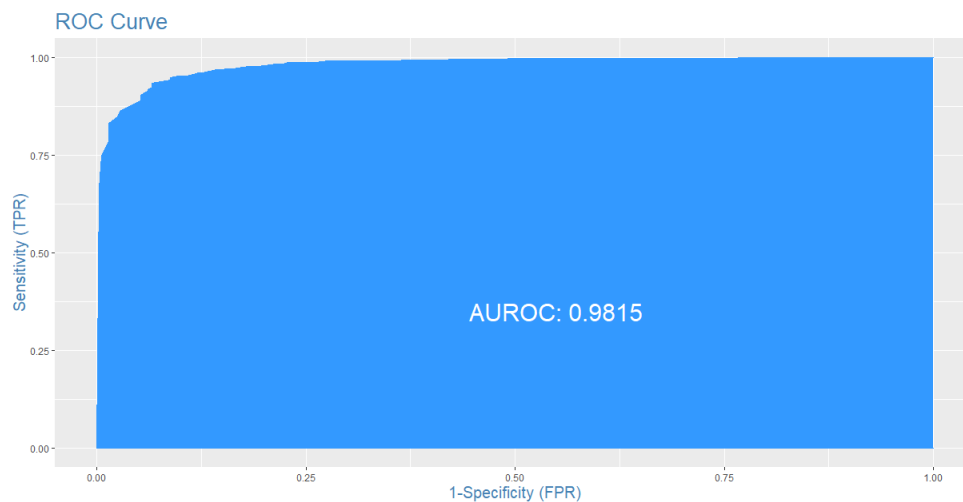
closer the negative correlation between the two variables is to -1, which means the negative correlation is greater. Through the above graph, we can compare the home wins and other variables in detail. It can be inferred from this that the offensive stability of home teams and the offensive flow achieved by assist are more influential on home wins. The influence of smoothness on home wins is the greatest. Also, in the figure we can find that FG_PCT_away and AST_away are negatively correlated to the greatest extent, i.e. the closer to -1. Here we can conclude that home teams also need to strengthen their defense as well as reduce the possible assists from the opposite team and interrupt the fluidity of the visiting team's offense, so as to minimize the probability of the opposite team winning.^v

Table 2: Logistic Regression for NBA Home Wins

| | Estimate | Std. Error | z value | Pr(> z) | |
|--------------|----------|------------|---------|----------|-----|
| (Intercept) | -0.594 | 1.838 | -0.323 | 0.747 | |
| FG_PCT_home | 11.090 | 1.471 | 7.538 | 0.000 | *** |
| FT_PCT_home | 3.734 | 1.026 | 3.639 | 0.000 | *** |
| FG3_PCT_home | 5.266 | 1.139 | 4.625 | 0.000 | *** |
| AST_home | 6.860 | 1.426 | 4.811 | 0.000 | *** |
| REB_home | 3.285 | 1.257 | 2.614 | 0.009 | ** |
| FG_PCT_away | -13.238 | 1.645 | -8.047 | 0.000 | *** |
| FT_PCT_away | -3.765 | 0.926 | -4.064 | 0.000 | *** |
| FG3_PCT_away | -4.504 | 1.266 | -3.559 | 0.000 | *** |
| AST_away | -6.099 | 1.233 | -4.947 | 0.000 | *** |
| REB_away | -3.149 | 1.185 | -2.657 | 0.008 | ** |
| sum_PF_home | -6.302 | 0.975 | -6.463 | 0.000 | *** |
| sum_PF_guest | 8.766 | 1.391 | 6.302 | 0.000 | *** |

According to the logistic regression, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away, sum_PF_home, sum_PF_guest were significant predictors of the winning of the home teams. ^v

Figure : ROC in testing sample for Logistic Regression



This graph shows the ROC graph to help us determine if the predictions made by our model are accurate and informative. When the area of this graph is less than 0.5, it means that the prediction is not as good as a random choice of two. If the area of this graph reaches 1, it means that the decision made by the prediction is definitely correct. The area under curve was 0.9815. The optional cutoff time is 0.52. The mis-classification error was 0.06. The sensitivity rate is about 95.4% and the specificity is 90.4%.

| Cut-off | sensitivity | specificity |
|---------|-------------|-------------|
| 0.4 | 96.7% | 94.0% |
| 0.6 | 86.5% | 92.3% |

4. Discussions

According to the logistic regression, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, FG_PCT_away, FT_PCT_away, FG3_PCT_away, AST_away, REB_away, sum_PF_home, sum_PF_guest were significant predictors of the winning of the home teams. The area under curve was 0.9815. The optional cutoff time is 0.52. The misclassification error was 0.06. The sensitivity rate is about 95.4% and the specificity is 90.4%.

In this study, we identified important of predictors of NBA Home Wins in the United States, for example, % of made shots and total of personal fouls. This tool will be very helpful to understand features determining NBA Home Wins and to maximize the likelihood to win. By looking at the numbers for shooting percentage and foul control, and by extrapolating them for winning at home, better foul control is a priority for every team under home court. There is clearly a lot of room for improvement and refinement in this analysis. The first is that the analysis of the relationship between each variables and the home wins is not very comprehensive without further in-depth analysis of the size of their correlation with home wins. Secondly, the study does not expand the connection between variables other than home-wins, which is also a limitation of this study. Among the pie charts analyzed above, it can be seen that there is a large negative correlation between the away team's shooting percentage and the number of our rebounds, and it is worthwhile to study this point in depth. By studying this part, we can understand how the rebounding protection affects the opponent's shooting percent.^{iv}

References

ⁱ<https://www.theringer.com/2021/6/1/22462636/what-happened-to-home-court-advantage>

ⁱⁱhttps://en.wikipedia.org/wiki/Home_advantage

ⁱⁱⁱ"Il fattore campo si prende la rivincita [Home advantage takes revenge]". repubblica.it (in Italian). September 19, 2001.

^{iv}<https://www.basketball-reference.com/>

^v <https://www.kaggle.com/nathanlauga/nba-games>.