Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/34/20240714

Research on NBA winning percentage prediction

Junyi Li[1,6], Yuntian Xu[2,7,11], Zizheng Xu[3,8], Luca Pan[4,9], Shubo Zhang[5,10]

[1]Finance Capital University of Economics and Business, Beijing, 100070, China,
[2]Math University of California Irvine, Los Angeles, 92697, United States,
[3]High school, Shawnigan Lake School, British Columbia, V0R 2W1, Canada,
[4]High School, Vanke Meisha Academy, Shenzhen, 518000, China,
[5]AP Program, Beijing National Day School, Beijing, 100039, China,

[6]2744987716@qq.com
[7]yuntiax@uci.edu
[8]Zizhengxu2004@gmail.com
[9]prj20050518@gmail.com
[10]Shubo.zhang06@icloud.com
[11]Corresponding author

Abstract. The National Basketball Association, also known as the NBA, is a professional basketball league composed of 30 professional teams in North America and one of the four major professional sports leagues in the United States. This research delves into predicting winning percentages in NBA games using a data-driven approach. It utilizes three years' worth of comprehensive game data from 30 NBA teams (2014-2016), including actual outcomes and expert predictions. The study aims to acquire valuable teamwork and data research skills while contributing to sports analytics. Data preparation involves cleaning and selecting recent seasons for model training. The prediction model utilizes an iterative algorithm to estimate team strength coefficients based on game results. Though the model shows promise, its accuracy falls slightly short compared to bookies' predictions. The research provides valuable insights for predicting NBA game outcomes and suggests avenues for improving the model's accuracy through additional variables

Keywords: Sport performance analysis, Winging prediction, Predictive model, Data preprocessing, Iterative algorithm

1. Introduction

The NBA is divided into the Eastern Conference and the Western Conference, with each league divided into three divisions, each consisting of five teams. After the end of each season and before the start of the next season, there will be an NBA draft; The NBA summer league, NBA preseason, and NBA regular season games usually start in October. In February, there was a special performance event called the NBA All-Star Game. After the NBA regular season ended, the Eastern and Western leagues advanced from the top eight to the playoffs, determining the Eastern and Western championships and advancing to the NBA Finals. The most outstanding player was awarded the Bill Russell Most Valuable Player

DOI: 10.54254/2753-8818/34/20240714

Award in the NBA Finals [1-5]. This research aims to delve into the prediction of winning percentages in the context of NBA games. The foundation of this study is built upon the comprehensive game data of 30 NBA teams over a three-year period from 2014 to 2016. This dataset not only encompasses the actual game outcomes but also incorporates the game predictions provided by professional forecasting teams. The amalgamation of real game results and expert forecasts offers a unique opportunity to rigorously assess the accuracy and effectiveness of the prediction model.

The primary objectives behind conducting this research are twofold. Firstly, it serves as an immersive experience in the realms of teamwork and data research, equipping us with valuable skills and insights that can be applied to future research endeavors or professional pursuits. Through a collaborative effort, we strive to embrace a data-driven approach to unravel the complexities of NBA game outcomes, thereby contributing to the broader field of sports analytics.

Secondly, this study caters to our intrinsic curiosity and passion for the NBA. By amalgamating our interests with data research, we endeavor to unravel the secrets that lie behind determining the best-performing team in the NBA. The teachings and guidance received from Professor Emerson have provided us with the necessary tools to analyze, interpret, and draw meaningful conclusions from the data at hand.

Combining our ardor for basketball with data analytics, we are steadfast in our pursuit of delivering compelling and conclusive results. By employing sophisticated modeling techniques and rigorous statistical analysis, we aspire to develop a prediction model that not only astutely forecasts winning percentages but also enriches our understanding of the underlying dynamics influencing team performances in the NBA. Ultimately, we endeavor to present findings that are persuasive, informative, and potentially transformative for the world of sports analytics.

2. Data Preparation

This report covers a dataset of NBA game results containing 7,868 rows and 10 columns. Each column records key information about the game such as date, season, teams, final score, home and away positions, number of overtime games, point-spread information, and over-under information. Among them, the final score plays an important role, by calculating the difference between the scores, we can get the points spread, and the average of the point spread will be the target variable (S_) for our model training. In the process of data cleaning, we first filtered out the files with a moderate amount of data and correct format from the given 18 CSV files, and finally chose the file named “6.csv” as our data source and carried out data cleaning.

The first step of data cleaning is to deal with duplicate rows, which reduces the size of the dataset to 55343 rows and 11 columns. Next, we addressed the issue of missing values NA value. Five missing values were found in the number of overtime games column, and by querying the NBA game records, we replaced these missing values with 0, 0, 0, 1, and 0 to complete the dataset.

In order to improve the prediction accuracy, we chose the three most recent seasons, i.e., 2014, 2016, and 2016, as the training data for our model. The reason for this is that these three seasons have the most match data, and there were no changes in the rules of the game and fewer changes in the players between these three seasons, which is conducive to improving the prediction accuracy on a team-by-team basis. The match data has been sorted by match date.

In summary, this report describes in detail the basic situation of the dataset and the data cleaning process, which lays the foundation for the subsequent model construction. Through reasonable data processing, we were able to better analyze and predict the data, explore the factors affecting the winning percentage of NBA games, and derive reliable and valuable insights.

3. Data Analysis

The iterative algorithm plays a key role in our prediction model, which is a method to optimize the objective function by constantly changing the values of the variables and is widely used in the field of machine learning. Here we use this approach to obtain strength coefficients for each team.

The script begins by initializing several features for each unique team in the dataset. These features include the strength coefficient (represented as "S_"), the number of games played (represented as "Pt_"), and the total score difference before the next game (represented as "Mdf_"). All these attributes are initially set to 0. Once the initialization is done, the script starts to iterate over all the games of different seasons, performing several operations for each game.

During each iteration, the strength coefficients from the previous game are carried forward to the current game, provided it's not the first game of the season. The script then predicts the score differential for the current game. If it's the first game for either team ("Pt_" for the team is 0), the prediction is computed as the difference between the teams' strength coefficients ("S_"). If it's not the first game, the prediction is the difference between the average score differentials ("Mdf_" divided by "Pt_") of the two teams. This prediction is further adjusted based on the game's location.

If the game is played at home, the prediction is reduced by 1.0. Conversely, if the game is played away, the prediction is increased by 1.0.

Subsequently, the script calculates the residual, defined as the difference between the actual score differential and the predicted score differential. It then updates the strength coefficients of the teams based on this residual. Specifically, the coefficient of team1 is reduced by the residual divided by 10, and the coefficient of team2 is increased by the same amount. In addition to updating the strength coefficients, the script increments the "Pt_" values for both teams, indicating that they have participated in one more game. It also updates the "Mdf_" values for both teams by adding or subtracting the actual score differential from their current "Mdf_" values. These iterative updates can give a prediction value and allow the algorithm to refine the team strength coefficients and improve prediction accuracy as it processes more games.

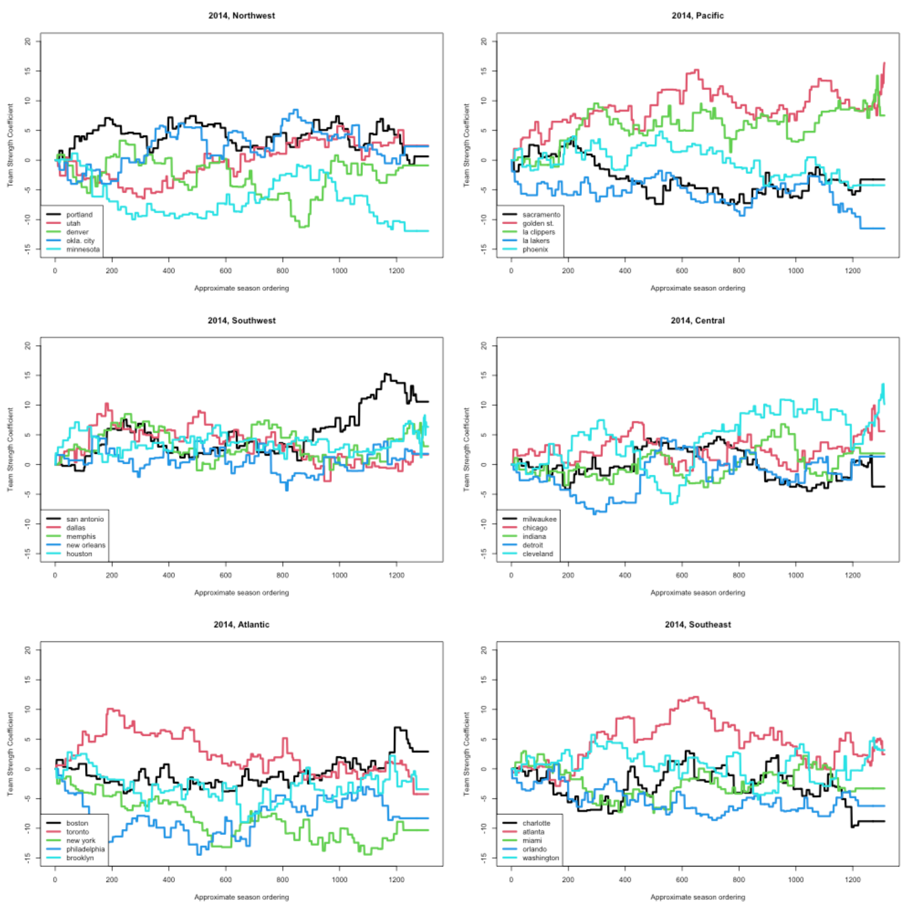

Figure 1. Strength coefficients for different teams over the course of the 2014 season

DOI: 10.54254/2753-8818/34/20240714

These six charts are divided based on the six divisions of the NBA. Figure 1 represents the change in strength coefficients for different teams over the course of the 2014 season. The strength coefficients, unique for each team, are a measure of the estimated strength of the team. The three teams with the highest strength coefficients were the Golden State Warriors, Cleveland Cavaliers, and Sanantonio Spurs, and that year's champion was the Sanantonio Spurs.

Figure 2. Strength coefficients for different teams over the course of the 2015 season

Figure 2 represents the change in strength coefficients for different teams over the course of the 2015 season. The three teams with the highest strength coefficients were the Golden State Warriors, Cleveland Cavaliers, and Sanantonio Spurs, and that year's champion was the Golden State Warriors.

DOI: 10.54254/2753-8818/34/20240714

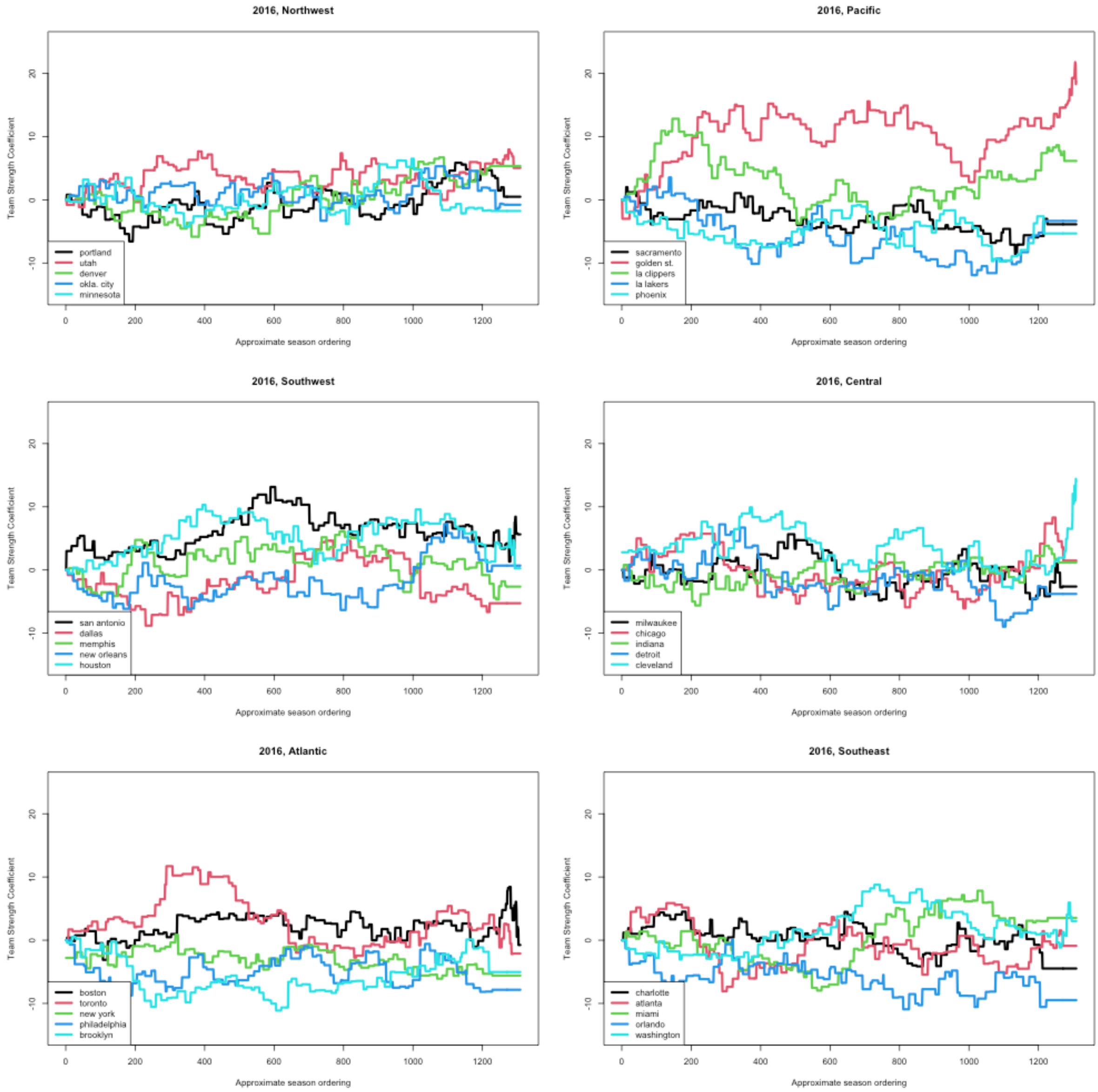

Figure 3. Strength coefficients for different teams over the course of the 2016 season

Figure 3 represents the change in strength coefficients for different teams over the course of the 2016 season. The three teams with the highest strength coefficients were the Golden State Warriors, Cleveland Cavaliers, and Boston Celtics, and that year's champion was the Cleveland Cavaliers.

By looking at Figure 1, Figure 2, and Figure 3, we can visualize the dynamics of the 30 teams' power coefficients, with higher coefficients resulting in higher win rates. The teams with the highest winning percentage in each season include the Golden State Warriors, Cleveland Cavaliers, San Antonio Spurs, and Oklahoma City Thunder, and it is from these teams that the Western Conference Champions, Eastern Conference Champions, and Overall Champions are born each season, which proves that the prediction has a certain degree of accuracy.

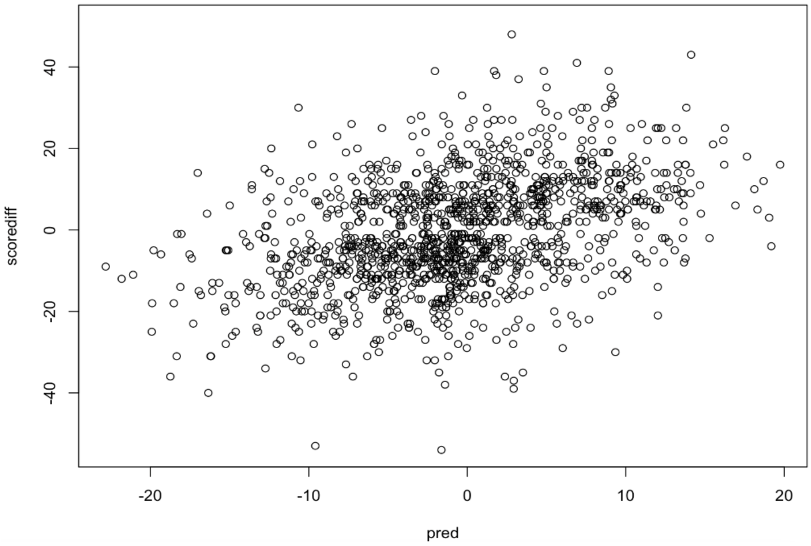

Figure 4. Relationship between the group's predicted results and the true score

This scatterplot shows our predicted (pred) and actual observed (observed) values for point spread for NBA games (Figure 4). point spread is a team's expected advantage or disadvantage in a game, expressed in points. This chart illustrates the following facts:

There is a slight positive correlation between predicted and observed values, meaning that when a team is predicted to have a greater advantage or disadvantage, the actual results tend to do the same. This suggests that there is some accuracy in your prediction methodology, but that accuracy is not forward to the point where you can accurately predict whether you will win or lose every game.

There is no clear pattern of deviation between our predicted and observed values, i.e., you do not systematically overestimate or underestimate a team. This suggests that your prediction methods are not obviously biased and ensure that the results are fair.

There is a bias between our predicted and observed values, meaning that we frequently overestimate or underestimate a team's performance. Although this does not have a very significant effect on the outcome of the tournament, it still suggests that our prediction methodology has errors and may need to be improved.

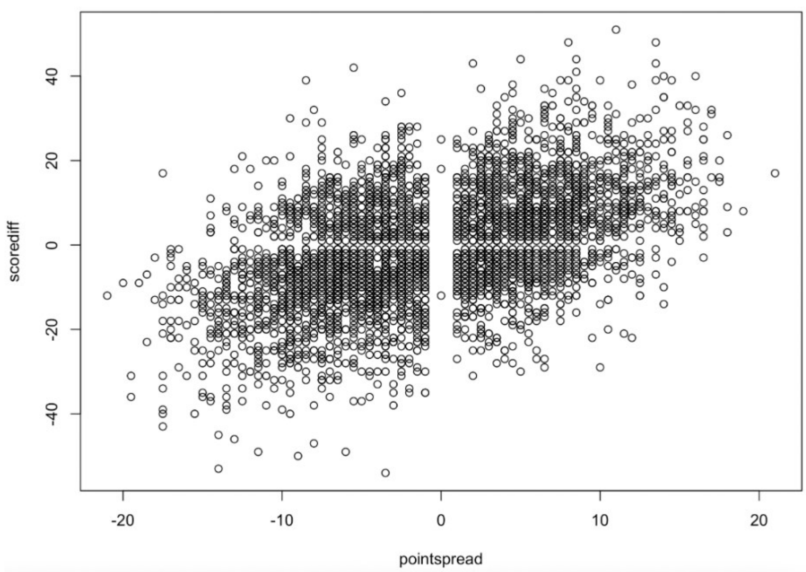

Figure 5. Relationship between the bookies' predicted results and the true score

This scatterplot shows bookies' prediction (point spread) and actual observed (observed) values for point spread for NBA games (Figure 5). Notice that there is a significant gap in the middle of this chart, which suggests that bookies rarely give predictions with a point spread of 0. That is to say, bookies hardly ever think that there are two teams that are evenly matched or that there are games that will end in a tie. This could be because the bookies want to attract more bettors, or because NBA games hardly ever end in a tie. This chart also illustrates the following facts:

There may be some bias or error in the bookie's prediction of point spreads for NBA games because the data points show a rectangular distribution in the graph rather than a normal distribution, i.e., the bookie may be inclined to give some extreme or unreasonable point spreads, It's possible that this is dictated by the speculative nature of the betting industry.

The bookies' predictions for point spreads for NBA games are not accurate because many of the data points are off the x-axis, meaning that the actual scoring differentials and the predicted scoring differentials have discrepancies. The predictions given by the bookies are also not very accurate.

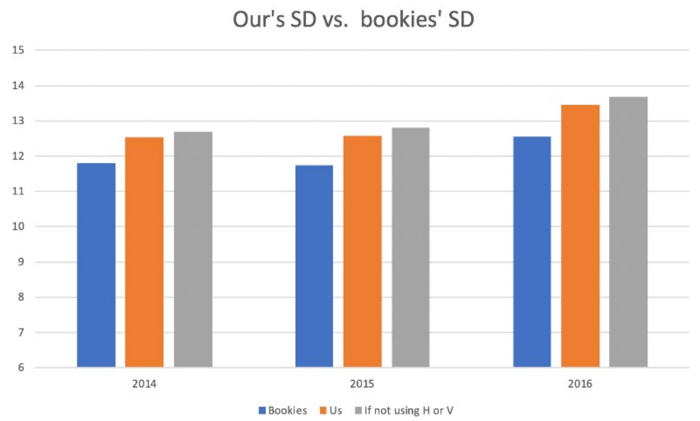

Figure 6. Comparison of the SD of our forecasting methodology with that of the bookies'

This bar chart compares the standard deviation of our forecasting methodology with that of the bookies from 2014 to 2016 (Figure 6). Standard deviation is a measure of forecast error, the smaller it is the better. By looking at this graph we can conclude that there is no significant difference in standard deviation between our forecasting method and the bookies' forecasting method, as the blue and orange bars are almost as high, meaning that our forecasting error is comparable to the bookies' forecasting error. Your prediction method uses the home and away factor (H or V), which has some effect on reducing the standard deviation because the gray columns are slightly higher than the blue and orange columns, which means that our prediction error would have increased a little bit if we didn't use the home and away factor. Considering that bookies are also often wrong in their predictions, it can be assumed that the error introduced by our prediction method is not significantly different from that of the bookies.

4. Conclusion

We created a predictive model to estimate the score differentials in NBA games. It involves calculating team strength coefficients based on game results and using those coefficients to make predictions for future games.

Our Group's NBA explorations present an experimental analysis aiming to predict score differentials in NBA games using a new approach. The analysis leverages historical game data to compute team strength coefficients, allowing for game-by-game updates. The predictions are made by considering the difference in coefficients between the competing teams. The experimental model's performance is compared against the baseline prediction made by bookies. The evaluation metric used to compare the predictions is the standard deviation of the differences between the predicted score differentials and the actual score differentials. While the experimental method shows promise, it still falls short when compared to the baseline prediction performance. However, it demonstrates a relatively reasonable level of accuracy. We believe that in future research we can increase the accuracy of the prediction results by adding more variables such as player turnover rate, on-court impact value, three-point specific statistics, and other variables to improve the model's imputation of team power coefficients.

References

[1] Silva J V R D, Rodrigues P. C. The three Eras of the NBA regular seasons: Historical trend and success factors[J]. 2021.DOI:10.3233/JSA-200525.
[2] Kai Liu, Zhuoya Li, Xinmin Li. NBA playoff first round prediction based on semi supervised clustering [J]. Data Mining, 2022, 12 (4): 10. DOI: 10.12677/HJDM. 2022.124030
[3] Jiangtao Zheng. Analysis of the Three Point Shooting Techniques of the Warriors in the 2016-2017 NBA playoffs [J]. Contemporary Sports, 2021 (6): 0108-0108
[4] Biyi Zhou and Fei Deng. Comparative study on height, weight, and age of athletes in the American NBA league and the Chinese CBA league [J]. Sports Leisure: Mass Sports, 2022 10): 0158-0160
[5] Zhihan Zhang, Zhixin Mao. Analysis of the Relationship between Technical Characteristics and Team Achievements of NBA Teams in the 2015-2016 Season [J]. Contemporary Sports Technology, 2021, 11 (30): 7