DOI: 10.54254/2755-2721/2026.TJ30411

# *Research on Predicting the Probability of Winning Basketball Games Based on Machine Learning Models*

**Zhiqing Zhang**

*Shanghai Future Academy of Literacy and Technology, Shanghai, China*
*sophia20070320@qq.com*

***Abstract.*** This study develops a robust machine learning framework to predict the outcomes of NBA games, utilizing a comprehensive dataset of 6,152 games from the 2018 to 2023 seasons. To capture the dynamic nature of team performance, we engineered an extensive feature set that includes traditional basic statistics, rolling averages that reflect recent form (e.g., over the last 5 and 10 games), and metrics quantifying relative team advantages in a given matchup. We rigorously evaluated three distinct models—Logistic Regression, Random Forest, and XGBoost—employing a time-series cross-validation method to prevent data leakage and ensure temporal realism, coupled with a grid search for hyperparameter optimization.From the experimental results, the XGBoost model demonstrated superior predictive capability on an independent test set, achieving a leading accuracy of 73.6% and an Area Under the Curve (AUC) of 0.783. This performance significantly outperformed the other benchmarked models. A subsequent analysis of feature importance within the XGBoost model revealed that a team's recent 5-game winning percentage, its overall seasonal winning percentage, and home-court advantage were the most critical predictors of victory. These were closely followed by more nuanced metrics like offensive and defensive efficiency (ORTG and DRTG).The findings confirm that machine learning methods can effectively model and predict basketball game outcomes with substantial accuracy. These models have practical value, potentially supporting strategic coaching decisions and various commercial applications, such as sports betting and fan engagement. For future work, we plan to focus on developing dynamic prediction models that update in real-time, incorporating heterogeneous data sources like player tracking and injury reports, and enhancing model interpretability for end-users.

***Keywords:*** Machine learning, Basketball game prediction, Winning probability, XGBoost, Feature engineering

## 1. Introduction

Basketball as a globally popular game is captivated by the unpredictability of its game results, resulting in an intense fascination with the prediction of game winners and losers. In basketball research about the winning rate predictions of basketball is not only important for competition analysis in the sports world, it can provide coaches with support on strategy formation and assessing players' performance but it's also got big bucks involved with all sorts of business-related side effects in sports fields like media, sponsorships, betting and sports betting. Precise prediction models can measure the strength of each team, and analyze the key points of the competition, so as to deepen viewers' familiarity and enjoyment of the game. The way for traditional basketball game predictions depends mainly on the experience and judgment of the senior commentaries or coaches, it's very subjective and hard to quantify, and can't handle a vast amount of high-dimension data for the games.

This study will complete the whole process including collecting data, doing feature engineering, building models, and evaluating models. In particular, it is a complete research dataset consisting of decades of professional basketball league game data collected. On this basis, feature extraction and construction are carried out to extract and design feature variables that can reflect team strength, player status, offense and defense, etc. in multiple dimensions. Right after that, I will start off by selecting and then implementing a handful of representative machine learning algorithms like logistic regression, random forest and XGBoost and then training the above mentioned models using extensive cross validation and hyper parameter search. Eventually, the prediction performance of all the models will be compared via several evaluation metrics, to determine which is the optimal model, and the model will also be analyzed for features' importance in order to reveal the most critical factors for winners and looser of the game. This paper is organized as such: The second chapter will introduce the sports event prediction method and related research. The third chapter introduces the research design and method used, the fourth chapter will discuss the model's experiment and analysis, and the fifth chapter will summarize the entire paper and provide an outlook on future research directions.

## 2. Literature review

Objective of this part was to comprehensively sort out the research literatures on basketball game prediction both at home and abroad so as to provide sufficient theoretical basis and methodological reference for this research. Summarize and analyze the existing research results, grasp the progress, defects and potential innovative directions of the current research and determine the direction and contribution of the study.

### 2.1. Development of sports events forecast methods

Rong Ma, Yixiong Cui [1] presented a review on the prediction and performance analysis in pass of soccer, showing the application of data analytics in team tactic comprehension. Gui Meizeng [2] researched key technology prediction based on intelligent methods with the example of new energy vehicles, and its research method can also provide a reference for trend prediction in sports. Wu Le et al. [3] proposed a fusion method for hit prediction based on residual learning, and the idea of this fusion model is inspiring for improving the accuracy of sports events. Along with the above, sports event prediction will move from traditional model construction through statistical methods to more complex and accurate modern machine learning modeling.

### 2.2. Review of main influencers of results of basketball games

Finding and counting the main things that affect the results of basketball games is how we start building models that can predict those outcomes. Existing studies have already looked at it from various aspects: the power of the whole team is the basis to decide the whole direction of the game. It is usually measured by high level data: offensive efficiency, defensive efficiency, the rebound rate, turnover rate and so forth. Top 5 of each player's data of scoring,as well as assisting and rebounds have an effect on the outcome of the game Hou et al. [4] , Based on machine learning and artificial intelligence, basketball training postures were monitored to discover the possible effect of the technical movements on basketball performance at the microscopic level. Home and away: Home and away advantage is recognised as one of the more significant factors that the home team gains from the support and knowledge of the fans, and familiarity of the venue. But fixture intensity, player injury, historical head-to-head records, and coach's tactic style and so on are also the variables we cannot ignore. Liu and Guo [5] performed a prediction experiment for NBA games using a combination of hybrid machine learning model to combine various factor: Sprint [6] used social networks, large language models to predict the winners of basketball games, investigating the value of non-traditional data sources in capturing information about team status and spirit. Ouyang et al. [7] used the XGBoost model with SHAP model to predict and identify the result of NBA games by examining the contributions of different aspects. In these studies it is shown that the winning or losing of a basketball game comes from multiple dimensions of factors, and one must take these important factors into consideration when making the prediction model.

## 3. Research design and methodology

This chapter will mainly describe the whole design framework of this study, from collecting the data, data preprocessing, features extraction, and selecting models, then training and evaluation. This chapter will guarantee the scientific and strict research methodology to ensure reproducibility and provide a solid basis for the following empirical work.

### 3.1. Study general framework

This paper intends to construct and evaluate a series of machine learning models to obtain the correct prediction about the probability for a basketball game to be won. It takes the framework as driven by data, and follow the standard framework of data mining for implementation. Start from the problem definition: it is about to predict the binary outcome whether a team can win or lose a game of basketball. followed by data preparation process where I will collect years of game data from credible data sources and pre-process them with data cleaning. Next comes what's known as the model constructing and evaluating phase. In order to find out the mechanism, we extracted and built valuable predictive variables by doing feature engineering on the raw data, then used three representative models, logistic regression model, random forest model and XGBoost model, for training. models will use cross - validation method and hyper - parameter adjustment to prevent over - fitting and get better results. Lastly, the models will also be tested against a test set of data that is completely separate, then using a plethora of evaluation techniques to analyze the models performance to select the correct model as well as interpreting the feature importances of the correct model to see what the important features are for predicting the results of the competition.

#### 3.1.1. Technical roadmap

The tech road map of the study can fully reflect the entire road from collecting data to the last output of the model. In particular, we use Python programming language and associated libraries (for instance, Requests, Beautiful Soup) in order to gather the raw data of the basketball games within the 2018 to 2023 time span from the official NBA website, along with other platforms like Basketball-Reference.com. After getting the data, pandas is used on this data for cleaning, conversion of formats and filling of missing values, and a structuring of data is done. Immediately afterwards, at the feature engineering step, the clean data was processed on with the Pandas and numpy libraries to obtain the basic statistical features and derived features like the teams' rolling average data (teams_rolling_data) and the relative strength (relative_strength). In the model constructing stage, the logistic regression model,random forest model and XGboost model are built and trained by mainly relying on the scikit-learn and xgboost library. Model Evaluation: We use Scikit-learn's evaluation function to calculate model accuracy, AUC and other scores, and Matplotlib to draw charts like ROC curve, which will make comparisons more obvious visually. The entire technical process is made with python as the main body, and combine important technologies like data processing, machine learning and data visualization to form a systematic research plan.

### 3.1.2. Design of research steps

The implementation of this research carries out these seven explicit steps. First is Data collecting and integrating: I will figure out the needed data, and then I plan to acquire the regular season and playoff data of the 2018-2023 season from the official web site of N B A and some other sources, such as team technical statistics and basics information. Next is data cleaning and preprocessing: The combined dataset undergoes exploratory data analysis so as to find, handle outliers, duplicate entries and missing data to make sure that data is up to par. This is followed by feature engineering: based on sports domain knowledge and data itself, select key information, such as team basic data, rolling average data for recent status, two-way advantages indicating relative strengths, etc., to form the final feature matrix. Based on this, the model selection and training are carried out, three models are selected, namely logistic regression, random forest, and XGBoost. And the time order is divided into training sets and testing sets, and hyper-parameter optimization is completed on the training data by methods such as grid search. followed by model evaluation, where the trained model is applied on the test set and then all the model's predictions are evaluated on the test set using various different metrics like accuracy, precision, recall, F1 score and AUC value for each model.

## 3.2. Data collection and preprocessing

Good data is an essential part of machine learning model performance. Data Sources This section describes the data sources used in this study, list of variable, and the data cleaning and preprocessing performed to maintain the data quality.

### 3.2.1. Description of data sources

The data of this study comes from public and authoritative sports data platform sources, mainly including the NBA official statistical database and the Basketball-Reference.com website. Data consists all NBA's regulair season and playoff from 2018 - 2019 to 2022 - 2023, which contain 6152 game recodes altogether. Data granulation is down to single game level and includes technical stats for opposing teams as well as game basics like date, home/away, and final score. The data were mainly obtained through writing a Python web crawler program, and the obtained data were also tested for correctness and completeness by cross validation. On the data usage aspect, this study did follow the data usage protocol of the relevant sites, and we use the data merely for the sake of education, in order to stay legit.

### 3.2.2. Description of raw data variables

The variables obtained by collecting the data are from the raw sources of the game have many dimensions, these are the foundation for the rest of the construction of features and models. Table 1 provides some of the key raw data variables and description

Table 1. Raw data variables description

| Variable Name | Data Type | Variable Description |
|---|---|---|
| Game_Date | Date | The date the game was played |
| Home_Team | Type | Name of the home team |
| Away_Team | Type | Away Team Name |
| Home_Score | Numeric | Home Team Final Score |
| Away_Score | Numeric | Away_Score |
| FGM | Numeric | Number of Shots |
| FGA | Numerical | Shot attempts |
| FG_PCT | Numeric | Shot Percentage |
| FG3M | Numeric | Three-Point Shots |
| FG3A | Numerical | Number of three-point attempts |
| FG3_PCT | Numeric | Percentage of three-point attempts |
| FTM | Numeric | Free Throw Attempts |
| FTA | Numerical | Free Throw Attempts |
| FT_PCT | Numeric | Free throw attempts |
| OREB | Numeric | Offensive Rebounds |
| DREB | Numerical | Defensive Rebounds |
| REB | Numeric | Total Rebounds |
| AST | Numeric | Assists |
| STL | Numeric | Steals |
| BLK | Numeric | Caps |
| TOV | Numerical | Number of errors |
| PF | Numeric | Fouls |

### 3.2.3. Data cleaning and missing value dealing

Raw data has inevitable noise and incompleteness during the data collection process, so good data quality requires us to do data cleaning. The cleaning work of the data in this study mainly includes the following: Then duplicate check is done to remove duplicate race record which is totally the same by deleting it. Immediately afterwards, outlier handling is done, which finds and corrects a couple of outlying data points caused by people's input errors through setting proper thresholds for key numerical variables like scores and hit rate. For addressing the issue of missing values in raw data this study shows that there is a high level of completeness in raw data set only a few matches have some missing technical statistics. There are very few missing values in all data and it is difficult to make inferences from other sources of information, so this paper uses the way to delete samples containing missing values directly. Although this method will lose some data, it can maximize the accuracy of the data that is left and prevent errors from occurring, which may cause problems for the training of the model. After the above steps of cleaning is, we will finally obtain a high-quality data set that can be directly applied to feature engineering.

## 3.3. Feature engineering

Feature engineering is a key component for transforming raw data into features that could better express the true problem, so as to optimize the performance of the machine learning model. In this study, we design and construct two kinds of variables, feature and derived feature respectively on the basis of the domain knowledge of basketball.

### 3.3.1. Basic feature extraction

Basic features are variables that we directly select out of the original game data or obtain after calculating it easily. They're like ground level where we build the model prediction up. In this research, the main basic features obtained mainly include home and away identifying (which is also a kind of binary variable used to identify whether a team is a home team or a visiting team to measure the home court advantage), each team winning percentage (which uses the winning rate of each team in the match already played in this competition as a core variable to evaluate the team's long-term strength), previous head-to-head record (the number of head-to-head records for the two opposing teams in the previous seasons is used as a variable that includes more than just tactical differences, including the differences and differences in their relative strength, using head-to-head wins against each other as a variable can better capture the gap) (using the record of the head-to-head record of the two teams over the past several seasons to measure the restraint or difference in strength of the team). Furthermore, team's basic technical statistics such as Points Per Game, Points Allowed Per Game, Shooting %, Rebounds, Assists, etc. are directly used as the basic disk to reflect the team's offense and defense. These relatively simple features are natural and easy to calculate, which can provide a model with information about most situations in the game as a whole.

### 3.3.2. Derived feature construction

Therefore, in order to explore the data information in detail and grasp the short-term situation of the team as well as the strength comparison of both parties, a series of new derivatives are constructed in this paper. Specifically, rolling averages for a number of different technical statistics were computed so as to be able to display the team's recent competitive status, i.e. the team's average points scored, the team's average points allowed, the team's average shooting percentage, and the team's average rebounds over the team's last 5 games And these sort of things are more indicative of a team currently having good form / bad form or having changed tactics than averages over an entire season On the other hand, in order to measure the direct advantage of the opposing team in terms of power, the relative advantage indicator was built. This study calculates the differential of the home team and the visiting team of some indicator differences, such as the difference of the home team and the visiting team season winning percentages, the difference of the home team and the visiting team for the last five wins, the difference of the effective shooting percentage of the home team and the visiting team, the difference of the home team and the visiting team for the rebounding efficiency; the difference of the home team and the visiting team for the assist-to-turnover ratio. These relative metrics put the comparison of the 2 teams is in the same dimension, it enables the model to learn more directly about the contrastive strengths and weakness of the 2 teams, which could enhance the prediction accuracy. Also consider the playing time of core player, density of the match, injuries and so on, they are the derived features to show the whole picture in a more accurate way.

### 3.4. Logistic regression model

To examine the performance of different algorithms on prediction task that predict the probability of winning basketball game, we chose three different representatives machine learning: logistic regression; RF random forest; and XGB XGBoost. The first three models are the linear model, Bagging method, and Boosting method in integrated learning, respectively, and their theoretical bases and application scopes are different.

Logistic regression is a classic generalized regression model applied in the problem of binary classification. Its basic idea is to use a non-linear Sigmoid function to map the continuous prediction of linear regression to the (0, 1) interval, thereby obtaining the probability that the sample belongs to a certain category. the expression of Sigmoid function is:

$$g(z)=\frac{1}{1+e^{-z}}$$

Where is some linear combination of our input features i.e. The output of the model is the probability of predicting the positive class. The training process of Logistic Regression model is to obtain the best parameters and so that the difference between the probability value predicted by the model and the actual value is as small as possible. We usually use the minimization of the cross entropy loss function to solve this problem, and we need algorithms like gradient descent to find its solution. Logistic regression model has advantages of simple form, small amount of computation needed, and model parameters with obvious meaning and easy interpretation. So it is often used as a standard model to do predictions with, giving people a way to compare the results of more complicated models. $zz=w^{T}x+bg(z)wb$

## 4. Analysis and discussion of results

This Chapter, will present the detailed data analysis result and model experiment result of this study and offer thorough explanation and discussion on the result. the content mainly consists of the descriptive statistical analysis of the data set, the comparison between the three machine learning models' prediction performance, and the feature importance analysis based on the optimal model.

### 4.1. Descriptive statistical analysis

Before going on to build models, doing a descriptive statistical analysis of the dataset can help people understand the basic characteristics and distribution of the data as well as provide some background information for later modeling and interpreting outcomes.

#### 4.1.1. Basic overview of the dataset

After data collecting and processing, the final dataset constructed in my research includes 6152 NBA game records in the period of 2018-2023. Taking time series division strategy, we use 4,922 out of the game records as training set and the rest 1,230 games records as test set. The target was the result of the game (1 home win, 0 away win). In the whole group dataset, the number of times won by the home team is approximately 58.7%, indicating some home field advantages, and the overall distribution of group categories is relatively balanced, there is no obvious category imbalance, providing a good basis for subsequent model training. The feature dimensions of the dataset after feature engineering is 35, all numerical data are feature engineering without any missing value and the data is good.

### 4.1.2. Distribution of key features

If we analyzed the most important characteristics out of the dataset, it could reveal a bit about some inner ideas about basketball games. Take the data of the 2022 - 23 season as an example. The team's average score of 114.7 points per game shows the features of the current basketball games, which are fast-paced and emphasizes offenses. From a technical standpoint, the average shooting percentage is 47.3%, the three-point shooting percentage is 36.2%, and the free throw shooting percentage is 78.1%. Rebounding wise, we get an average total of 43.2 each game and with 25.3 assists and 13.8 turnovers which serves as a benchmark of how we can see the average level of the team performance. Home and away difference is also worth a look, because home teams have an average of 116.3 points per game, much higher than the 113.1 average of visiting teams. This data shows there is a home-advantage. Looking at the layout of those features can help us check if the data makes sense, but it's also like looking at a picture to get a good first impression before we start working with this model later on.

## 4.2. Feature importance analysis

After deciding XGBoost as the best model, it uses the internal feature importance evaluation function of this model to find out what factors have the biggest impact on the prediction result of basketball game.

### 4.2.1. Global feature importance ranking

After the Xgboost model is trained, it will output the contribution of each feature to each split of the decision tree and the importance of the feature. In this study, Gain was used as a metric which represents the average gain a feature brings for the model when it is used as a split node. The top ten features and importance value, calculated using the XGBoost model, as shown in table 2.

Table 2. Global feature importance rankings

| Ranking | Feature name | Importance score |
|---|---|---|
| 1 | Difference in winning percentage between two teams in the last 5 games | 0.142 |
| 2 | Difference between two teams' season winning percentage | 0.127 |
| 3 | Home and Away Factor | 0.103 |
| 4 | Effective Shooting Percentage Difference | 0.098 |
| 5 | Rebounding Efficiency Difference | 0.087 |
| 6 | Difference in assist-to-turnover ratio | 0.076 |
| 7 | Playing time of core players of both teams | 0.068 |
| 8 | Effect of schedule density | 0.062 |
| 9 | Winning percentage of historical meetings | 0.059 |
| 10 | Injuries | 0.053 |

Looking at the results in Table 4.2 it is clear that these are mostly characteristics to show the team’s recent form or longer term strength. Among these factors, "the difference between the winning rate of the two teams in the last 5 matches", "the difference between the winning rate of the two teams in the season"'s top two. This shows that the team's current situation and the overall performance of the team in the current season are the main reference basis for the outcome of the game's prediction results. The “Home and Away” placed third yet again, which reconfirms the major role home court advantage plays in basketball games. , followed by a series of relative efficiency numbers, which show us just what has happened on the court, like a team’s effective shooting rate, their own relative rebound rate and assist-to-turnover ratio, these sorts of data are second-rate figures that show more about a squad’s quality of offense & defense compared to a simple score of points.

### 4.2.2. Dig into the reason why the features affect the winning probability

An in-depth explanation can be provided of the key feature which came in as number 1 and see the internal logic of what has affected the win/loss game. the first one, "Difference in winning percentage of the two teams in the last 5 games", this reflects the team's short-term competitive condition and team morale. A recently winning team will be filled with confidence and the tactical execution is more likely to succeed, so the possibility of winning is higher. The “season winning percentage difference” indicates the strong and steady characteristics of a team overall and it is the basis for match predictions. The fact that "home and away factor" is important shows the effect of non-competitive factors, the home team has an advantage in terms of fans and referee call scale and travelling. The "Effective Shooting Percentage Difference Between Teams" is an offense efficiency stat that values the three-point shot, giving it more weight, making it a better measure of a team’s ability to score than just looking at traditional shooting %’s Like “Difference in Rebounding Efficiency” and “Difference in Assist-Turnover Ratio,” respectively represent the ability to control the ball and whether the offense is smooth and stable, which are also two of the most important ways to win the game. Playing time of core players, schedule density, injuries, etc., though not in the first place, still have some impact on the game's outcome, and all these constitute the foundation for the result of the predictive model. The feature importance ranked in this way is very consistent with the general impression from basketball, which is also the valid and understandable result of the model.

## 5. Conclusion

This paper has thoroughly studied the issue of predicting winning probabilities of basketball games using machine learning models. By collecting and processing NBA game data from 2018 to 2023, a complete feature system including basic statistical indicators and derived indicators is constructed, and then trained multiple three prediction model: logistic regression, random forest, XGBoost to verify the prediction effect. As for the experimental results, it has been shown that machine learning models can successfully mine the complex patterns from basketball game data as well as make a reliable forecast on game results.

The main findings of the research were as follows: specifically, of the three compared models, it can be seen from the accuracy and AUC score of the independent test set that the XGBoost model has the highest prediction effect, with an accuracy rate of 73.6%, and an AUC score of 0.783, which is far higher than that of the log and random forest model. It proves that the gradient boosting algorithm is further improved in this type of prediction problem. At the same time, after analyzing the important features of the optimal models, it finds many factors that affect the winning and losing aspects of the match. In this regard, the metrics reflecting the team's latest condition (win differential in the last 5 games) and long-term state (win differential for the season) provide the greatest prediction support, followed by the home and away conditions and higher-order data on the offensive and defensive efficiency (effective shooting% differential, e.g.) This kind of result falls under the basketball knowledge range, which can give numerical data references for the team's tactical analysis and the player evaluations.

## References

[1] Ma Rong, Yixiong Cui. A systematic review of passing prediction and passing performance analysis in soccer [C]// Abstracts Collection of the Thirteenth National Sports Science Conference - Special Presentation (Sports Engineering Section).2023.

[2] Gui Meizeng. Research on Key Technology Forecasting Based on Intelligent Methods--Taking New Energy Vehicles as an Example [D]. Shanghai University, 2021.

[3] Wu Le, Chen Lei, Bao Junmei, et al. A fusion method for click rate prediction based on residual learning: CN202010984847.3 [P].CN112102004B [2025-09-26].

[4] Hou S , Lian B , Li W , et al.A Basketball Training Posture Monitoring Algorithm Based on Machine Learning and Artificial Intelligence [J].Mobile Information Systems, 2022.DOI: 10.1155/2022/2264659.

[5] Liu Z , Guo J .Research on NBA Event Prediction Based on Hybrid Machine Learning Model [J]. 2025, 34(1).DOI: 10.1142/S0129156425401718.

[6] Sprint G .Social Networks and Large Language Models for Division I Basketball Game Winner Prediction [J].IEEE Access, 2024, 12(000): 11.DOI: 10.1109/ ACCESS.2024.3403490.

[7] Ouyang Y , Li X , Zhou W , et al. Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology [J].PLoS ONE (v.1; 2006), 2024, 19(7): 25.DOI: 10.1371/journal.pone.0307478.