# Using Machine Learning Models to Predict Win Rate of NBA Teams

Jieyu Han

High school affiliated to Renmin University of China, Beijing, 10080, China

**Abstract.** NBA is the biggest sports organizations, the behavior of teams and the players in the teams have a big influence on the economic developments. When selecting different players for the team, the coach needs to consider about the data and the thorough behavior of players. Big data and machine learning provide this question a more precise and proper solutions. In this paper, a forecasting method of player's influence on team's winning rate based on technical statistics of players. Experiment shows that the proposed method is effective.

**Keywords:** NBA; Winning Rate; Predict; KNN.

## 1. Introduction

In the highly competitive realm of the NBA, every team is in pursuit of an edge, and understanding how individual players impact team performance is key to gaining that advantage. Traditional statistical methods, such as weighted averages and efficiency ratings, have been employed to measure player performance, but these approaches often fail to capture all the intricate factors that influence game outcomes. With the rise of big data and machine learning, new methodologies have begun to emerge, offering more nuanced and dynamic assessments of player impact.

Existing literature includes examples of using machine learning for sports analytics, yet these studies often focus on short-term performances or specific technical statistics. There remains a scarcity of research focusing on the long-term prediction of a player's impact on team success through machine learning. Moreover, there is a lack of consideration for how a player's injuries, psychological state, and interactions with teammates can affect overall team performance.

This study aims to use advanced machine learning algorithms, such as Random Forest and Neural Networks, to predict the future impact of draft players on team performance by considering a comprehensive set of statistics from before joining the NBA and during their first three years in the league, alongside data on the players' physical conditions. Through an in-depth analysis of historical data, we expect to reveal more intricate correlations between player performance and team success. Not only will this analysis provide guidance for team management, aiding them in making more informed decisions, but it will also offer feedback to players and coaches, helping them to better understand and enhance their contributions to the team.

We anticipate that this research will offer a new perspective on data analysis and decision-making for NBA teams, and will be of relevance to any sports organization looking to improve athletic performance through data science.

## 2. Related Works

There is currently a lot of research on utilizing machine learning algorithms for the analysis and prediction of sports games.

[1] discusses how machine learning algorithms are used to predict the outcomes of football matches, including methods for data preprocessing, feature engineering, and model selection.

In [2], Sandeep Kumar utilizes machine learning techniques to predict the outcomes of cricket matches and compares the performance of different algorithms, including decision trees, random forests, and support vector machines. [3] provides an overview of various machine learning techniques used in predicting the outcomes of soccer matches and analyzes their pros and cons as well as performance. The research in [4] explores how machine learning algorithms are employed to

predict NBA game results and discusses issues such as feature selection and model evaluation. The study in [5] attempts to predict the champion of the 2014 FIFA World Cup using machine learning methods and addresses challenges related to dataset construction and feature engineering.

[6] reviews the use of machine learning to predict soccer player performance, including passing, scoring, and defensive aspects. The survey [7] reviews the applications of machine learning in various sports including soccer, basketball, tennis, and discusses related data mining and feature engineering techniques. The research in [8] employs machine learning techniques to predict the outcomes of professional golf tournaments and discusses strategies for feature engineering and model evaluation. The study in [9] explores how machine learning is used to predict the performance of basketball players, including scoring, assists, and rebounds. The survey in [10] provides an extensive overview of the applications of machine learning in sports analytics, including predicting match outcomes, analyzing player performance, and optimizing game strategies.

## 3. Our Method

From the perspective of machine learning, the decision to select a player is actually based on predicting whether that player will contribute to the team's future performance. To achieve this goal, this paper proposed a method that relies on typical technical statistical data of players before the draft and in their first three years in the NBA to forecast how a player is likely to impact the team's win rate over the next five years.

In this method, the attributes utilized comprises a player's technical statistics, such as score, assist, and rebounds in his first season, and these attributes are then used to construct our training and testing datasets.

Our method consists of three steps: 1) data collection; 2) Model selection; 3) KNN-based prediction modeling. The following sections will descript these steps in detail.

### 3.1 Data Collection

The data used in this paper is collected from the Official Website of NBA. We collect different players technique data from the website, meanwhile, we also use relative website to gain the data before the players are selected. The beneath form is a kind of form to get players' data. The previous data of the player is mainly collected from the player's data in the NCAA match or the other state competitions.

**Table 1.** The Sample Data Of The Player In NBA

| name | score | assistance | backboard | rate | Successful rate of his team |
|---|---|---|---|---|---|
| James Harden | 23.6 | 4.7 | 2.4 | 1.4 | 68.7 |

**Table 2.** The sample data of the player before he steps in NBA

| name | score | assistance | backboard | rate | Successful rate of his team |
|---|---|---|---|---|---|
| James Harden | 18.7 | 5.2 | 2.6 | 1.7 | 63.5 |

### 3.2 Model Selection

In theory, all regression-based machine learning or deep learning models can be used for predicting the contribution of NBA players to team win rates. However, due to the specific characteristics of NBA data and certain aspects of the prediction scenario itself, different algorithms and models may exhibit variations in performance when applied to the same forecasting task. Therefore, for the purposes of this study, it is essential to carefully select the algorithm or model to be used.

Based on the work in Section 3.1, due to the highly dispersed nature of the data, collecting technical statistical data for NBA players is a challenging task, making it difficult to gather a substantial amount of data for building predictive models. Furthermore, as a premier sports league, the number of players that can join the NBA each year is very limited, with the majority of players entering the league

through the draft, which typically admits around 60 players annually. Therefore, fundamentally speaking, there won't be a large pool of sample data available for use.

Due to the limited amount of data, algorithms and models that heavily rely on a large dataset are not ideal choices. Additionally, drawing inspiration from the evaluation methods used in the NBA, player assessment typically involves selecting template players from the pool of NBA players based on their characteristics and assessing the future development of a target player by considering the development trajectory of these template players.

In summary, this paper opts for the K-Nearest Neighbors (KNN) algorithm for the prediction task, primarily for two reasons: (1) KNN aligns more closely with the concept of evaluating draft prospects based on template players, reflecting the fundamental approach used in NBA player assessment; (2) KNN is better suited for prediction tasks in scenarios with limited data samples.

### 3.3 The Variable-Weighted KNN Model

#### 3.3.1 Traditional KNN Model

The K-Nearest Neighbors is a classical machine learning model. This method is distributed to two tasks, the classification and aggression.

The core principle of this algorithm is to determine the label of a test sample based on its similarity to labeled samples. To enhance the algorithm's generalization capability, it introduces a hyperparameter K, which decides the test sample's label by taking a vote from the K most similar samples. The basic principle is illustrated in the following diagram.

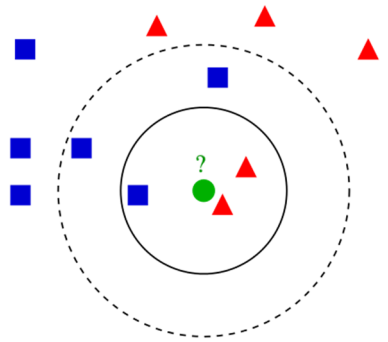


**Figure 1.** The principle of KNN algorithm

#### 3.3.2 The Variable-Weighted KNN

Since we cannot just use data to reflect the rate. Players in the court have different position while playing, which means the data he contributes to the team also different. For example, taking Stephan Curry as an example, it is really hard to find the relationship between Curry's backboard and the team winning rate, because Curry is not for doing that. Besides, the score and the assist are the common things among player, since it both have huge contribution to the winning rate.

Considering the varying impact of different attributes on evaluating a player's contribution to the team, the distance calculation for comparing players, as opposed to traditional similarity calculation methods, is as follows:

$$d = |X - Y| = \sqrt{(x_1 - y_1) \cdot p_1 + (x_2 - y_2) \cdot p_2 + ... + (x_n - y_n) \cdot p_n}$$

Where $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ are two samples, $P = (p_1, p_2, ..., p_n)$ denotes the attribute weight vector, and $n$ is the number of attributes for each data object. Based on this idea, the classical KNN algorithm is improved and called VW-KNN (Variable-Weighted KNN).

Regarding the distance formula in this paper, it is essential to specify how to allocate appropriate weights to each attribute. Intuitively, attributes that exhibit a stronger correlation with the target prediction attribute should have a more significant influence when calculating similarity. Following this idea, this paper proposes a method for weight allocation based on correlation coefficients. In this paper, Pearson's coefficient is used as the measure of correlation strength. After statistical analysis,

the correlations between various player performance metrics and team win rates are presented in Table 3.

**Table 3.** The correlation between the statistics indexes of player and the winning percentage of the team

| statistics index of player | Correlation |
|---|---|
| Score | 0.883 |
| Assist | 0.375 |
| Rate | 0.430 |
| Backboard | 0.693 |

Based on the derived correlations between statistics index of player and team win rates, the weights for each technical performance attribute are calculated using the following formula.

$$p_i = \frac{c_i}{\sum_{i=1}^{K} c_i}$$

Where $c_i$ represents the correlation of the ith statistics index of player, and K denotes the number of statistics indexes. According to the formula above, the weights for each technical performance attribute are as shown in Table 4.

**Table 4.** The weight of each statistics indexes

| Statistics index of player | Weight |
|---|---|
| Score | 1.126 |
| Assist | 2.663 |
| Rate | 2.322 |
| Backboard | 1.423 |

### 3.4 Experiments

This section tests the performance of the proposed variable-weight K-Nearest Neighbors (KNN) algorithm using the NBA player technical statistics dataset obtained in Section 3.1. The dataset comprises a total of 100 data points, which are divided into a training set and a test set, with 70 data points in the training set and 30 data points in the test set.

To demonstrate the effectiveness of the proposed algorithm, the experiments also include testing various other algorithms such as Decision Tree, Linear Regression, RandomForest, AdaBoost, Bagging, Gradient Boosting, Extra Tree, and SVR. In these tests, the percentage of team wins over the next three years is used as the target value for prediction, with Root Mean Square Error (RMSE) serving as the evaluation metric for prediction accuracy. The test results are summarized in the table 5.

**Table 5.** The result of experiment

| Algorithms | RMSE |
|---|---|
| Decision Tree | 24.08 |
| Linear Regression | 15.88 |
| Random Forest | 16.18 |
| AdaBoost | 15.74 |
| Bagging | 17.44 |
| Gradient Boosting | 17.75 |
| Extra Tree | 15.9 |
| SVR | 14.59 |
| VW-KNN | 13.65 |

The test results indicate that the proposed VW-KNN algorithm in this paper exhibits the best performance with an RMSE of 13.65, which is lower than all the compared algorithms. The primary reasons for this are as follows: (1) VW-KNN, built upon the K-Nearest Neighbors (KNN) framework, aligns more closely with the concept of using template players to evaluate prospective NBA players, making it more suitable for scenarios with limited data samples; (2) VW-KNN takes into consideration the varying impact of different technical statistics on player evaluations, thus better reflecting the influence of different player attributes on their future development.

## 4. Conclusion

This article proposes an enhanced K-Nearest Neighbors (KNN) algorithm for predicting the future contributions of NBA draft prospects to team win rates based on their technical statistical information. This algorithm considers the varying impact of different technical statistics on a player's contribution, thus incorporating the idea of assigning different weights to different technical statistics when calculating player similarity. Experimental results show that, when using team win rate as the prediction metric, the proposed algorithm outperforms the comparison algorithms in terms of prediction accuracy.

## References

[1] S. Vijayalakshmi, R. S. D. Wahidabanu, "Predicting Football Matches Results using Machine Learning Techniques", International Journal of Computer Applications, 2011.

[2] Sandeep Kumar, et al. "A Machine Learning Approach for Predicting Cricket Match Outcome", International Journal of Computer Applications, 2015.

[3] Eliezer Silva, et al. "A Survey of Machine Learning Techniques in Predicting the Outcomes of Soccer Matches", International Journal of Computer Applications, 2013.

[4] Joshua F. P. Koo, et al. "Predicting NBA Game Results Using Machine Learning Techniques", Proceedings of the 2015 International Conference on Data Mining.

[5] Gregory J. Matthews, et al. "Using Machine Learning to Predict the 2014 FIFA World Cup Champion", Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems.

[6] Lev Muchnik, et al. "Predicting Player Performance in Soccer: A Comprehensive Review", Journal of Sports Analytics and Data Science, 2020.

[7] Xinyu Zhang, et al. "Machine Learning in Sports Analytics: A Survey", IEEE Transactions on Neural Networks and Learning Systems, 2020.

[8] Daniel P. Palomar, et al. "Using Machine Learning to Predict the Outcomes of Professional Golf Tournaments", Proceedings of the 2020 IEEE International Conference on Data Mining Workshops.

[9] Jiaxin Zhang, et al. "Predicting the Performance of Basketball Players: A Machine Learning Approach", Proceedings of the 2017 IEEE International Conference on Data Mining Workshops.

[10] Aniket Bera, et al. "Machine Learning for Sports Analytics: A Comprehensive Survey", ACM Computing Surveys, 2020.