Machine Learning Approaches to Forecasting the Winner of the 2024 NBA Championship

Hana Zadravec
hana.zadravec@student.um.si
University of Maribor
Faculty of Electrical Engineering and Computer Science
Maribor, Slovenia

ABSTRACT

Forecasting the winner of the NBA Championship has become more important as there is a large amount of data and the league's popularity is increasing. This research investigates techniques in machine learning to predict the winner of the 2024 NBA Championship. Three methods - random forest regression, SVR, and linear regression - are used and assessed. The process includes scraping data from Basketball Reference, then analyzing and selecting features. Findings show the leading projected teams for 2024 according to each model, with random forest regression showing the best prediction. Analysis of feature importance emphasizes critical predictors like team quality rating and player performance metrics. The research highlights the capabilities of machine learning in predicting sports outcomes and indicates areas for additional research to improve predictions.

KEYWORDS

forecasting, basketball prediction, statistical analysis, NBA Championship, machine learning

1 INTRODUCTION

Forecasting the winner of the NBA championship has become increasingly accessible for sports analysts, bettors, and enthusiasts alike. This endeavor prompts the exploration and application of sophisticated analytical methodologies to enhance predictive precision. The necessity for more precise prognostications is underscored by recognizing the NBA's status as the most extensively followed professional sports league in 2022, engaging 2.49 billion individuals [4]. Comprising 30 teams in North America, the NBA stands as a premier basketball league showcasing elite players globally [5]. With an annual revenue surpassing $10 billion, the league continuously accumulates a wealth of data crucial for analysts and strategic planning within sports organizations seeking competitive advantages through data analysis. This data often informs pivotal on-field decisions regarding team formations and gameplay strategies, such as offensive or defensive approaches. Such insights can significantly impact match outcomes. Moreover, this wealth of data facilitates individual game outcome predictions in the realm of NBA contests. Ahead of each match, numerous analysts proffer their forecasts for the victor. These predictions are scrutinized by commentators on NBA platforms who provide pre-game analysis. Furthermore, a growing betting industry has arisen around prognosticating NBA matchups. This sector expands annually with a key emphasis on developing precise models adept at handling pertinent metrics in NBA games effectively. Hence, the increasing integration of machine learning models in sports represents a pivotal and adaptable strategy moving forward.

The research motivation comes from the necessity to improve sports analytics in the NBA, aiming for more precise predictions to benefit strategic decisions and operations in the betting sector. Due to the constraints of current models that often overlook important metrics, this research aims to enhance prediction accuracy by utilizing different machine-learning techniques. This study also seeks to address a deficiency in the literature, as it seldom focuses on predicting the champion of the entire championship.

This article examines forecasting NBA championship winners by utilizing three machine learning techniques: random forest, support vector regression (SVR), and linear regression. Section 2 will examine pertinent studies in the field of predicting sports performance, specifically honing in on NBA results. In Section 3, we provide a comprehensive explanation of the techniques utilized for gathering and examining data, as well as the implementation of the specified models. Next, we will discuss the results and evaluate how well each technique worked in Section 4. Section 5 explores the importance of our discoveries for future research in this area. Finally, our analysis leads to conclusions in Section 6.

2 LITERATURE REVIEW

Advancements in predictive modeling for basketball are increasingly important within sports analytics. With the growing integration of machine learning in this field, researchers continue to explore strategies for improving predictions. This section reviews studies focused on forecasting basketball outcomes using various machine-learning techniques.

Yongjun et al. [3] propose using data analysis to forecast NBA team performance by combining statistical regression methods to predict the relationship between game results and winning chances. They apply Data Envelopment Analysis (DEA) to identify optimal performance standards, evaluating their process with the Golden State Warriors, which demonstrated high predictive accuracy. The study suggests enhancing predictions by incorporating rival tactics and expanding the model for game-level forecasts for each player.

Bunker and Susnjak [1] analyze the use of machine learning methods for forecasting match outcomes in team sports. They evaluate various algorithms, including regression models, decision trees, and neural networks, assessing their predictive success with past data and player statistics. The study emphasizes advancements in machine learning that enhance accuracy and the challenges faced in

practical applications. It also highlights the importance of data quality and feature selection in improving sports analytics, providing valuable insights for future research.

Yao [8] assesses how well neural networks predict outcomes compared to traditional regression models using past NBA data. The results indicate that regression models provide straightforward explanations, while neural networks are better at capturing intricate patterns, leading to increased precision. This research highlights how advanced machine learning methods can be utilized in sports analysis to improve performance prediction strategies.

The study by Huang and Lin [2] introduces an innovative method for predicting game results using regression tree models. The writers examine different elements impacting game results, like player data and team interactions, to create a targeted predictive system for the Golden State Warriors. Their research shows that regression trees can capture intricate connections in the data, leading to precise predictions of scores.

Thabtah et al. [7] explore the use of machine learning to forecast NBA game outcomes in their research. Numerous learning models, such as decision trees, artificial neural networks, and Naive Bayes, have been applied. After analyzing the data, it was discovered that important characteristics including total rebounds, defensive rebounds, three-point percentage, and the quantity of made free throws are essential for accurately predicting the outcome of games.

3 METHODOLOGY

In this section, we provide a comprehensive explanation of the approach utilized for examining NBA data in our research. All analyses were performed using a Jupyter notebook. The primary objective was to collect, process, and analyze NBA data to develop models for forecasting outcomes for the 2024 NBA season.

3.1 Data Collection

The initial step involved gathering data through web scraping methods. We sourced data from the Basketball Reference website [6], which offers extensive statistics for every NBA season up to the present day.

Web scraping was chosen for its efficiency in collecting large volumes of data without manual input. We used Python libraries such as BeautifulSoup and requests to retrieve HTML content from the web pages. The pertinent data, including player stats, team stats, and game outcomes, were extracted and organized into CSV files for ease of manipulation in subsequent analyses.

3.2 Data Preprocessing

Following data collection, we performed meticulous processing to ensure data quality and consistency. This included:

- We filled missing NaN values with the median to preserve the dataset and minimize their impact on the analysis.
- Standardizing the data as needed to maintain uniformity across the dataset.
- Normalizing features to bring them into a common scale, which is especially important for algorithms sensitive to the magnitude of input features.
- Encoding categorical variables into numerical format using label encoding, allowing for their inclusion in machine learning models.
- Removing duplicates to eliminate redundancy in the dataset and improve model performance.

3.3 Feature Selection

To enhance model performance, we addressed multicollinearity by filtering features based on their correlation. Pearson's correlation coefficient was used to assess the linear relationship between features. The coefficient r is calculated as:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

where $\text{cov}(X, Y)$ is the covariance between variables X and Y, and σ_X and σ_Y are their standard deviations. Pearson's r ranges from -1 to 1:

- $r = 1$ indicates a perfect positive linear relationship,
- $r = -1$ indicates a perfect negative linear relationship,
- $r = 0$ indicates no linear relationship.

We set a correlation threshold of 0.9, identifying features with an absolute value of Pearson's r above this threshold as highly correlated. Features exhibiting high correlation were removed to reduce redundancy and mitigate multicollinearity, thus enhancing model interpretability and reliability.

Additionally, we utilized Principal Component Analysis (PCA) as a dimensionality reduction technique to transform the feature space, allowing us to reduce the number of features while retaining essential information, further enhancing model performance.

3.4 Model Selection and Data Splitting

Data was divided into training and testing sets as follows:

- **Training Data:** Data from the 2005 to 2023 NBA seasons.
- **Testing Data:** Data from the 2024 NBA season.

We employed three machine learning models to forecast NBA game outcomes:

3.4.1 Support Vector Regression (SVR). **Reason for Selection:** SVR is selected for its capability to handle complex, non-linear relationships between features and the target variable (game outcome). SVR is effective in scenarios where interactions between variables are intricate.

Advantages: SVR finds optimal hyperplanes to minimize prediction errors within a specified margin, capturing subtle patterns in the data.

3.4.2 Random Forest Regression. **Reason for Selection:** Random Forest regression was selected for its ensemble method, combining decision trees to improve predictions and manage overfitting. It effectively captures complex interactions in NBA data.

Advantages: This model handles high-dimensional data, identifies key features, and is robust against outliers. It also models non-linear relationships with minimal tuning, making it versatile for both categorical and continuous variables in sports analytics.

3.4.3 Linear Regression. **Reason for Selection:** Linear Regression is chosen as a baseline model for predicting NBA champions due

to its simplicity and interpretability, clarifying how features affect winning likelihood.

Advantages: It offers easy interpretation of feature impacts, with coefficients showing expected outcome changes for unit shifts in predictors. Minimal computational resources are needed for quick training and evaluation, and it serves as a reference for comparing more complex models.

3.5 Experiment

Our experiment focuses on NBA data from the 2005 season onward, due to significant changes in gameplay and statistical tracking. Prior to 2005, the game was more physical and lacked modern statistics like three-point shooting (3P%), which were introduced post-1990.

We used data from the 2005 to 2023 NBA seasons to train our models, which included 49 features related to team and player performance. Some of the features are:

- `pre_season_odds`: The odds assigned to each team before the season starts, indicating their chances of winning the championship.
- `team_rating_custom`: A custom rating for each team based on various performance metrics, reflecting their overall strength.
- `FG%`: Field Goal Percentage, representing the ratio of field goals made to field goals attempted, a key indicator of shooting efficiency.
- `3P%`: Three-Point Percentage, indicating the ratio of three-point field goals made to three-point attempts, measuring a team's effectiveness from beyond the arc.
- `max_player_rating_custom`: A custom rating for the highest-rated player on each team, capturing the impact of star players.

The target variable for prediction is `champion_share`, which represents the chances of a team winning the NBA Championship in the 2024 season. This continuous variable can take values between 0 and 1, where a higher value indicates a greater probability of a team being crowned champion. In the dataset, known winners from previous seasons are marked with a value of 1.0, signifying their championship status, while teams that did not win are represented by lower values closer to 0.

3.5.1 Model Parameters. Default parameters were used for all models:

- **SVR:** The Support Vector Regression (SVR) employs a Radial Basis Function (RBF) kernel, which is effective for capturing non-linear relationships. The regularization parameter $C = 1$ controls the trade-off between achieving a low training error and a low testing error, while gamma $\gamma = 0.1$ determines the influence of individual training examples on the decision boundary.
- **Random Forest:** This model consists of 100 trees with no maximum depth specified, allowing each tree to grow fully. This approach enhances model accuracy by averaging predictions from multiple trees, reducing the risk of overfitting.
- **Linear Regression:** The linear regression model uses default parameters, applying ordinary least squares to estimate coefficients without regularization. This simplicity allows it to fit the data by minimizing the residual sum of squares, serving as a benchmark for more complex models.

3.5.2 Evaluation Metrics. Model performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors. It is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

  where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the number of observations.
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

  where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the number of observations.

4 RESULTS

The results section provides a comparison of the models based on MAE and MSE metrics. Table and figure illustrate the performance of each model and highlight predictions for the top teams in the 2024 NBA season.

4.1 Model Comparison

Figure 1 presents the comparison of MSE and MAE across the three models used in our study. The Random Forest model achieved the lowest MSE and MAE, indicating superior performance in predicting NBA outcomes compared to SVR and Logistic Regression.

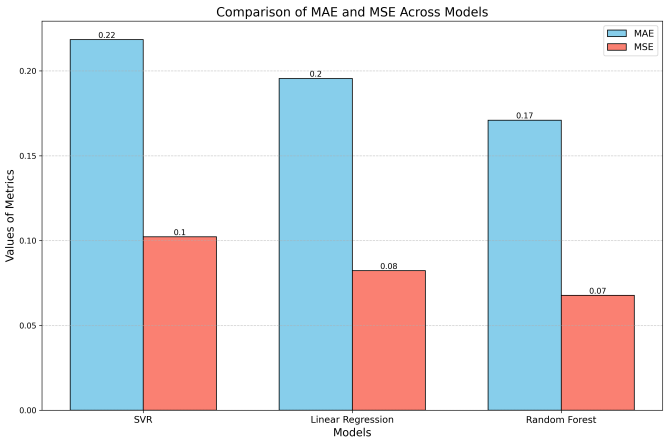

Figure 1: Comparison of MSE and MAE for each model.

4.2 Model Predictions

Table 1 presents a summary of the predicted top teams for the 2024 NBA Championship as determined by each model, along with their corresponding `champion_share` values. These values indicate the estimated probability, expressed as a decimal, of each team winning the NBA Championship in the 2024 season.

Table 1: Top 3 Predicted Teams for 2024 NBA Championship

Model	Top Predicted Teams	Predicted `champion_share`
SVR	Milwaukee Bucks	0.8691
	Boston Celtics	0.7510
	Denver Nuggets	0.6726
Random Forest	Boston Celtics	0.6485
	Milwaukee Bucks	0.5643
	Minnesota Timberwolves	0.5527
Linear Regression	Denver Nuggets	0.6706
	Milwaukee Bucks	0.6448
	Boston Celtics	0.6227

5 DISCUSSION

In this study, we assessed the performance of three predictive models—Random Forest, SVR, and Linear Regression—regarding the NBA Championship outcome for the 2024 season. The results reveal several insights and distinctions between these models.

5.1 Random Forest Model

The Random Forest model achieved the lowest MSE and MAE in predicting the NBA Championship outcome. Its ability to capture complex feature interactions enabled it to identify key determinants of the championship. Notably, it predicted the Boston Celtics as the 2024 season winners, aligning with the actual outcome and validating its predictive performance.

5.2 Support Vector Regression

Although the SVR model did not perform as well as the Random Forest and Logistic Regression models in terms of predictive metrics, it was effective in revealing intricate relationships between features. The SVR model assigned high predicted probabilities to both the Milwaukee Bucks and Boston Celtics, reflecting their strong performances throughout the season. However, the actual season outcome highlighted significant challenges for the Milwaukee Bucks, such as injuries to key players and a mid-season coaching change. These factors likely impacted their final standing, demonstrating that while SVR provided valuable insights, it may not fully account for unforeseen disruptions and their effects.

5.3 Linear Regression

Linear Regression, while not as effective as the Random Forest model in predictive metrics, still provided valuable insights. The model's predictions for the Boston Celtics and Denver Nuggets as strong contenders aligned with the final outcome of the championship. This highlights the model's utility in scenarios where more complex methods might be less interpretable. Despite its limitations, Linear Regression contributed to a broader understanding of potential championship winners.

5.4 Real-World Outcome

At the end of the 2024 season, the Boston Celtics were confirmed as the champions, validating the Random Forest model's prediction and partially supporting the SVR model's forecasts. Despite high values assigned to the Milwaukee Bucks by the SVR model, their performance was hindered by significant issues such as player injuries and a coaching change, which affected their final standing. The Minnesota Timberwolves, who were also predicted to be in contention, remained competitive until the end of the season, demonstrating that our models were accurate in predicting some outcomes.

6 CONCLUSION

This research assessed various machine learning techniques in forecasting the 2024 NBA season such as SVR, Linear Regression, and Random Forest models. The Random Forest model outperformed others, showing its capability to deal with intricate feature relationships by achieving the lowest MSE and MAE. Even though SVR and Logistic Regression were not as accurate, they still offered important information on team performance, highlighting the difficulties encountered by the Milwaukee Bucks because of injuries and coaching adjustments. This study highlights the significance of forecasting the whole season instead of single games.

One noteworthy aspect of this study is the emphasis on making predictions for the entire season, as opposed to just individual games. Our results indicate the importance of integrating current data and regularly updating it to enhance the accuracy of predictions. Future research should aim to integrate real-time data with advanced modeling techniques to more effectively adapt to the dynamic conditions and changes that occur throughout the NBA season.

In summary, incorporating various machine learning models and adjusting predictions with real-time data can improve the precision of sports predictions.

REFERENCES

[1] Rory Bunker and Teo Sušnjak. The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 75:1–22, 2022.

[2] Mei-Ling Huang and Yi-Jung Lin. Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry*, 12(5):835, 2020.

[3] Y. Li, L. Wang, and F. Li. A data-driven prediction approach for sports team performance and its application to national basketball association. *Omega*, page 102123, 2019.

[4] National Basketball Association. About the nba. https://www.nba.com/news/about, 2024. Accessed: 2024-04-30.

[5] PlayToday. Nba viewership statistics. https://playtoday.co/blog/stats/nba-viewership-statistics/, 2024. Accessed: 2024-04-30.

[6] Basketball Reference. Basketball reference. https://www.basketball-reference.com/, 2024. Accessed: 2024-04-30.

[7] Fadi Thabtah, Ling Zhang, and Nadia Abdelhamid. Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116, 2019.

[8] Alan Yao. Comparing neural and regression models to predict nba team records. *Frontiers in Artificial Intelligence and Applications*, 320:421–428, 2019.