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

Hana Zadravec
hana.zadravec@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 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, SVR, and logistic regression - are used and assessed. The process includes scraping data from Basketball Reference, then analyzing and extracting features. Findings show the leading projected teams for 2024 according to each model, with Random Forests showing the best precision. 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 accuracy in forecasting.

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 logistic 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

Continuous advancements in improving predictive accuracy in sports such as basketball are essential within sports analytics research. With the increasing utilization of machine learning technology in this domain, researchers are exploring innovative strategies to improve predictive models. This section will focus on machine learning techniques for forecasting the outcomes of basketball games.

Houde [1] developed a machine-learning model to enhance current models. By utilizing Python and machine learning methods, he examined team data, incorporating two extra characteristics. He tried six different models, and Gauss' Naive Bayes produced the top outcomes. He performed an extensive search on the network to enhance the model, which had an average accuracy rate of 65.1% in forecasting the results of NBA matches.

In his article, Lieder [2] created a model using machine learning techniques that considers a range of NBA game features. The model's ultimate accuracy reached nearly 70%. When developing the model, no information regarding team injuries or lineup before the game was considered. The writer emphasizes how this information could help enhance the precision of his model.

Lin et al. [3] created a model to predict NBA game winners using data from 1991 to 1998. Based on the data, the Random Forest model accurately predicted the outcome 65.15% of the time. Through the division of the season into quartiles, logistic regression resulted in a 68.75% increase in accuracy. In most cases, eliminating teams' latest wins resulted in a 2-3% reduction in prediction accuracy.

Wang [8] uses machine learning methods such as logistic regression, support vectors, KNN, and random forests to predict NBA game outcomes in his study. Predictions are formulated based on team statistics and individual player's performance. He determined that the random forests and KNN models give the best forecasts for NBA game results, emphasizing the importance of the team's shooting percentage.

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:

- Removing rows with missing (NaN) values.
- Standardizing the data as needed to maintain uniformity across the dataset.

3.3 Feature Extraction

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.

3.4 Model Selection and Data Splitting

Data was divided into training and testing sets as follows:

- **Training Data:** Data from the 1990 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. **Reason for Selection:** Random Forest is chosen for its ensemble approach, combining multiple decision trees to enhance prediction accuracy and manage overfitting. It is well-suited for the diverse features in NBA data.

Advantages: Random Forest handles high-dimensional data effectively and provides insights into feature importance.

3.4.3 Linear Regression. **Reason for Selection:** Linear Regression serves as a baseline model due to its simplicity and interpretability. It models the linear relationship between features and game outcomes.

Advantages: Provides a straightforward interpretation of feature impacts on outcomes, serving as a reference for more complex models.

3.5 Experiment

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

We used data from the 1990 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`, representing the likelihood of a team winning the NBA Championship in the 2024 season.

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

- **SVR:** Radial Basis Function (RBF) kernel, regularization parameter $C = 1$, and gamma $\gamma = 0.1$.
- **Random Forest:** 100 trees with no maximum depth specified.
- **Logistic Regression:** L2 regularization with a regularization strength parameter $C = 1.0$.

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.
- **Accuracy:** Measures the proportion of correctly classified instances out of the total number of instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

  where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.
- **Precision:** Measures the proportion of true positive instances among all instances classified as positive. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

  where TP is the number of true positives and FP is the number of false positives.

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 2 presents the comparison of accuracy and precision across the three models used in our study. The Random Forest model demonstrated the highest accuracy and precision, outperforming both Linear Regression and SVR. This indicates that Random Forest provides the most reliable and consistent predictions among the models evaluated.

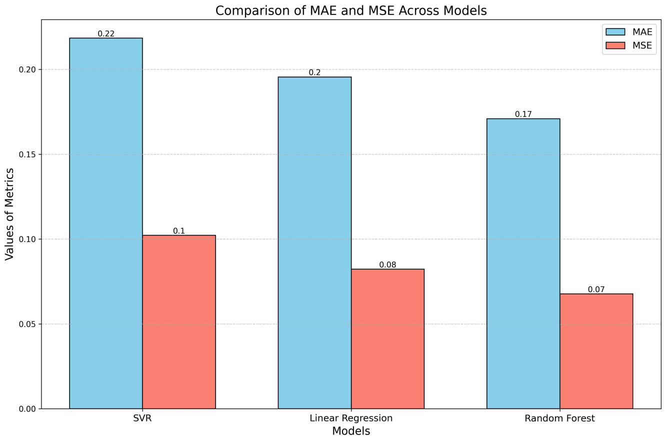

Figure 1: Comparison of MSE and MAE for each model.

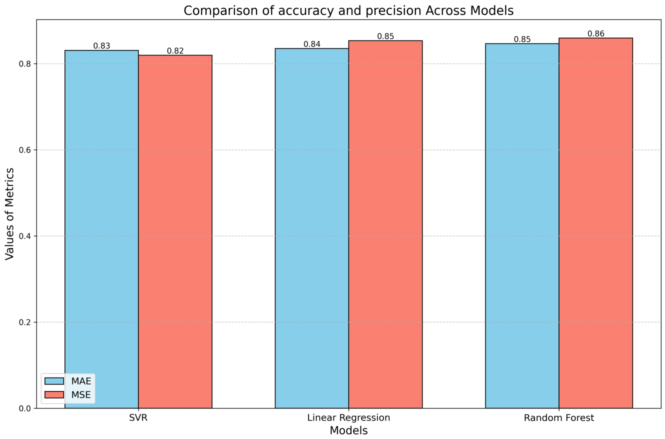

Figure 2: Comparison of accuracy and precision for each model.

4.2 Model Predictions

Table 1 summarizes the predicted top teams for the 2024 NBA Championship by each model, including their predicted probabilities. These probabilities represent the estimated chance, expressed as a percentage, 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 Probability
SVR	Milwaukee Bucks, Boston Celtics, Denver Nuggets	86.91%, 75.10%, 67.26%
Random Forest	Boston Celtics, Milwaukee Bucks, Minnesota Timberwolves	64.85%, 56.43%, 55.27%
Logistic Regression	Denver Nuggets, Milwaukee Bucks, Boston Celtics	67.06%, 64.48%, 62.27%

5 DISCUSSION

In this study, we assessed the performance of three predictive models—Random Forest, SVR, and Logistic 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 demonstrated the highest accuracy in predicting the NBA Championship outcome, as indicated by its lower MSE and MAE. This model's ability to capture complex interactions between features and accurately identify the key determinants of the championship outcome is reflected in its superior performance. Specifically, the Random Forest model predicted the Boston Celtics as the winners of the 2024 season, which aligns with the actual outcome, affirming the model's predictive accuracy.

5.2 Support Vector Regression

Although the SVR model did not achieve the same level of precision as the Random Forest and Logistic Regression model, 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 exposed significant challenges for the Milwaukee Bucks, including injuries to key players and a mid-season coaching change. These factors likely affected their final standing, demonstrating that while SVR provided useful predictions, it may not fully account for unforeseen disruptions and their impacts.

5.3 Logistic Regression

Logistic Regression, while less accurate compared to Random Forest model, still offered valuable insights. The model's predictions for the Boston Celtics and Denver Nuggets as strong contenders were consistent with the final outcome of the championship. This highlights the model's utility in scenarios where other methods might be less effective. Despite its lower precision, Logistic 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 probabilities 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, Logistic 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] M. Houde. Predicting the outcome of nba games. *Bryant Digital Repository, Honors Projects in Data Science Senior Honors Projects*, 4, 2021.

[2] Nachi Lieder. Can machine-learning methods predict the outcome of an nba game? March 1 2018.

[3] Jason Lin, Lance Short, and Vinay Sundaresan. Predicting national basketball association winners, 2014. CS 229 Final Project, Autumn 2014.

[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] Junwen Wang. Predictive analysis of nba game outcomes through machine learning. In *Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence (MLMI '23)*, pages 46–55, New York, NY, USA, 2024. Association for Computing Machinery.