NBA Game Outcome Predictor Using Machine Learning

Omar Elsegeiny
omar.elsegeiny@ndsu.edu
North Dakota State University
Fargo, North Dakota, USA

Xander Johnson
xander.johnson@ndsu.edu
North Dakota State University
Fargo, North Dakota, USA

Figure 1: Phoenix Suns against the Brooklyn Nets, January 2024

Abstract

Predicting the outcome of NBA games has become an important task for sports analysts, bettors, and fans alike. With advances in machine learning and access to large datasets, it is now possible to create predictive models that forecast game results based on historical data and performance metrics. This paper presents an approach to predict the outcome of NBA games using machine learning techniques, using data sourced from the NBA API. Various classification models are explored, including logistic regression, support vector machines, ridge classifiers. Data processing, feature engineering, and model evaluation are central to the study. The findings suggest that machine learning models can provide valuable information on the prediction of game outcomes, with potential applications in sports analytics and betting.

Keywords

NBA, Machine Learning, Predictive Analytics, Classification Models, Sports Betting, Data Science

1 Introduction

The prediction of NBA game outcomes has been a topic of interest for various stakeholders, including sports analysts, teams, coaches, fans, and bettors. Traditionally, predictions were based on expert opinions and analysis, but with the advent of machine learning and the increasing availability of historical and real-time data, predictive models have gained prominence. By analyzing key player statistics, team performance metrics, and advanced analytics, these models aim to predict the probability of a win or loss for each team. This paper aims to develop a predictive model for the outcomes of NBA games using machine learning techniques. The main objective is to leverage the NBA API, a rich source of player and team data, to build a robust classifier that can predict the outcome of games. The application of this model extends beyond sports analytics, with potential benefits for bettors and fans who seek data-driven insights into game results.

2 Data Collection

The primary data source for this study is the NBA API, which provides a wide range of historical and real-time data related to NBA games. The following types of data will be utilized:

- Historical Game Results: Outcome of previous games (win or loss), box scores, and more.
- Player Statistics: Individual player statistics such as points, assists, rebounds, steals, and blocks, that add up to team totals.
- Team Performance Metrics: Team-level statistics, including plus/minus ratings, and field goal percentages.

Additional data sources to enhance the model may include:
(for future implementation)

- Betting Odds: For baseline comparisons, betting odds from sports books will be analyzed to compare with the predictions generated by the model.
- External Data: Information on player injuries, team schedules, and other factors influencing performance.

3 Methodology

3.1 Data Pre-processing

Data preprocessing is a crucial first step in any machine learning project. The following steps will be undertaken to prepare the data for modeling:

- Data Extraction: Extract relevant data using the NBA API.
- Cleaning and Transformation: Handle missing values, normalize features, and convert categorical variables to numerical representations.
- Feature Engineering: Create new variables such as rolling averages of player and team win/loss statistics.
- Standard Scaler: A preprocessing technique that standardizes features by removing the mean and scaling to unit variance, ensuring all variables contribute equally to the model.
- Data Splitting: The dataset will be split into training and testing sets, to train the models and evaluate their performance.

3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis will help identify trends and correlations in the data. Statistical techniques and visualizations will be used to explore relationships between variables, such as:

- Correlations between team performance metrics and game outcomes.
- Distribution of winning and losing teams based on different features.

3.3 Machine Learning Models

The goal is to predict the categorical outcome of NBA games (win or loss). Several machine learning classification models will be applied:

- Logistic Regression: A simple baseline model to assess the relationship between predictors and the binary outcome (win/loss).
- Ridge Classifier: A linear model that applies L2 regularization to reduce overfitting and improve generalization, especially when predictors are highly correlated.
- Time Series Split: A cross-validation strategy that respects the temporal order of data by training on past observations and validating on future ones, ideal for time-dependent datasets.

Each model will be trained on the dataset, evaluated using appropriate metrics, and optimized for hyper parameters to ensure the best performance.

4 Results

The evaluation of the machine learning models will rely on key performance metrics to assess their effectiveness in predicting NBA game outcomes. Among these, accuracy will serve as the primary metric. Accuracy in this instance is the proportion of correct predictions made by the model, reflecting its overall ability to distinguish between winning and losing outcomes. Accuracy provides a straightforward and interpretable measure of model performance, particularly useful in binary classification tasks like predicting win/loss results. To ensure the reliability and robustness of the evaluation, time-aware cross-validation techniques will be employed. This approach helps prevent data leakage and ensures that the model performs well on unseen, temporally ordered data. This is an essential consideration in sports prediction tasks.

5 Discussion

Logistic Regression and Ridge Classification will be used to gain insights into the factors influencing NBA game outcomes. These linear models are well-suited for binary classification tasks and offer interpretability, allowing for a clearer understanding of how different features contribute to win or loss predictions. Logistic Regression serves as a simple and effective baseline, while Ridge Classification incorporates L2 regularization to address multicollinearity and improve generalization. In addition, external factors such as player injuries and team schedule strength could significantly impact predictions. Incorporating such variables may further enhance the accuracy and reliability of the results.

6 Conclusion

This paper outlines a methodology for predicting NBA game outcomes using machine learning techniques. By leveraging data from the NBA API and applying various classification models, we aim to develop a tool that offers predictive insights for sports analysts, bettors, and fans. The findings suggest that machine learning models can successfully predict game outcomes, with potential applications in sports analytics, team performance evaluation, and betting.

Future work could focus on refining the model, incorporating more features, and exploring real-time data streams to enhance prediction accuracy. Furthermore, user interfaces and integration with other platforms could make the model more accessible and impactful for a wider audience.

Acknowledgments

This project is dedicated to our families, who we would not be here without.

A Data Source

- Swar, S. (2021). NBA API Documentation. Retrieved from https://github.com/swar/nba_api

B Research Sources

- Dataquest. (n.d.). Dataquest YouTube Channel. Retrieved from https://www.youtube.com/@Dataquestio
- Raschka, S., Mirjalili, V. (2019). Python Machine Learning (3rd ed.). Packt Publishing. ISBN: 9781789958294