NBA Sports Betting

Bet Big, Win Bigger: NBA Action Awaits!

Rahul Aneja
Computer Science
Virginia Tech
Blacksburg VA, United States
rahulaneja@vt.edu

Anton Bilonog
Computer Science
Virginia Tech
Blacksburg VA, United States
antonb@vt.edu

Gio Romero
Computer Science
Virginia Tech
Blacksburg VA, United States
gioromeroruiz@vt.edu

Figure 1: NBA Logo

ABSTRACT

Our project is about NBA sports betting and the primary goal is to examine and understand the machine learning strategies used by a group in their Spring 2023 titled "*Machine Learning in Sports Betting (NBA)*". With sports betting becoming an increasingly popular trend, there is significant potential for it to be lucrative over time for those who apply the right strategies.

Our primary goal is to critically analyze the methodologies employed, assess the reproducibility of the results using the latest models, and explore potential improvements. The broader aim is to contribute to the evolving field of sports analytics by enhancing the predictive capabilities of machine learning models in sports betting, thereby helping bettors make more informed decisions and potentially increase their profitability. This study not only seeks to validate earlier findings but also aims to refine the models to adapt to changes in data and betting landscapes in the ongoing NBA seasons.

1 Introduction

In the rapidly changing field of sports betting, particularly within the NBA, the proliferation of data has presented both challenges and opportunities for predictive analysis. The project "Machine Learning in Sports Betting (NBA)" directly addresses the complexities involved in forecasting game outcomes – an endeavor where precision can significantly enhance profitability. This project initiated in Spring 2023, employs advanced machine learning techniques to analyze patterns and predict NBA game results with oddly high accuracy.

The main project tackled by this project is the inefficiency and inaccuracy in predicting NBA game outcomes within the sports betting community.The technique involves the application of machine learning models to improve the prediction accuracy of NBA game winners. This approach not only seeks to outperform traditional predictive methods that are often based on simplistic statistical metrics or expert opinions but also to automate and streamline the betting process. Our approach used the most recent models from last year's project to predict ongoing 2024 bets and games to challenge the original results that were measured.

Our analysis focuses one key property of the employed machine learning models which is the

- **Coverage and Accuracy:** The project claimed that the model could cover 60% of NBA games while predicting 70% of the winners. This aspect was critically evaluated to determine the model's effectiveness across different game scenarios and seasons

We wanted to see if these models would hold up with the 2023-2024 NBA season and see if they obtained these high predictive accuracies. We addressed this property by tweaking their old models and using newer models with different parameters to adjust for relevant factors that could influence the game outcomes.

By conducting a comprehensive re-evaluation of the machine learning models with updated algorithms and parameters, our aim was to not only validate the previous seasons success but to also establish a robust framework that can continuously adapt to the evolving nature of NBA statistics and betting markets.

2 Related Work

There have been many different sorts of techniques and strategies to build machine learning models with the intention of making profit and beating the sportsbooks. An example we will be talking about is a study titled, "*A Study on Sentiment Analysis Techniques of Twitter Data*" which dived into the various methodologies employed for sentiment analysis, which particularly focused on Twitter data. It provides a comprehensive review of techniques such as supervised and unsupervised learning, lexicon-based approaches, and hybrid models. The study categorizes these techniques based on their application at different levels—document, sentence, and aspect/feature level. It also explores the use of machine learning classifiers like Naive Bayes, Maximum Entropy, and Support Vector Machines in analyzing sentiments expressed in tweets. This article's exploration of sentiment analysis techniques can be analogously related to our project. While the primary focus in our project is not on sentiment analysis, the underlying machine learning techniques and data handling strategies discussed in the articles are relevant. For instance, similar to sentiment classification, our project involves classifying and predicting outcomes (win/loss) based on historical and real-time data. Techniques such as feature extraction, data preprocessing, and the use of various machine learning algorithms discussed can inform the enhancement and optimization of predictive models in the NBA sports betting context. Additionally, the review of hybrid methods that combine different machine learning approaches could inspire innovative ways to integrate multiple predictive features or models to improve the accuracy and reliability of betting predictions. This cross-application of techniques highlights the versatility and adaptability of machine learning methodologies across different domains, including sports analytics and sentiment analysis. [1]

Another study titled, "*Exploiting Sports-Betting Market Using Machine Learning*" introduces a sophisticated system designed to generate profits from sports betting through advanced machine learning techniques. It presents three innovative approaches: reducing the model's correlation with bookmaker predictions, utilizing convolutional neural networks (CNNs) for match outcome prediction based on player statistics, and applying modern portfolio theory to optimize betting strategies. These techniques aim to improve profit generation by leveraging more accurate and less correlated predictive models. This article is highly relevant to our project as it focuses on similar goals of enhancing predictive accuracy and profitability in sports betting using machine learning. The use of CNNs to process detailed player and team statistics informed us with the approach to help build our models. [3]

In addition the article, "*Sports Data Mining Technology Used in Basketball Outcome Prediction*" discusses the use of machine learning algorithms for predicting NBA game outcomes. The research utilized various algorithms including Simple Logistics Classifier, Artificial Neural Networks, SVM, and Naïve Bayes to analyze data from five regular NBA seasons for model training and one season for testing. The study achieved the highest accuracy with the Simple Logistics Classifier at 69.67%, emphasizing the effectiveness of applying substantial datasets in predictive models. This article directly correlated with our project as it aligns with the project's focus on employing machine learning models to predict NBA game outcomes. The methodologies and results provided us a valuable comparative benchmark for assessing our predictive accuracy of the models used in our project. After reading this article this is where we drew the inspiration to use a Logistic Regression model in our methodology where they proved that it was the best for binary outcomes. [2]

The article "*A Machine Learning Framework for Sport Result Prediction*" presents an in-depth examination of machine learning (ML) techniques specifically for predicting sports results, emphasizing neural networks. It reviews the literature on the effectiveness of artificial neural networks (ANNs) in sports prediction, noting past successes and areas needing improvement. The paper proposes a

CRISP-DM type framework specifically tailored for sport result predictions, highlighting the need for comprehensive domain understanding, meticulous data preparation, and rigorous model evaluation to improve prediction accuracies. Furthermore, the article's discussion on the challenges of model overfitting and the importance of feature selection resonates with our need to build robust models that do not merely fit historical data but are also predictive of future games. The insights into training and testing methodologies, especially the recommendation against using cross-validation due to the sequential nature of sports data, will directly influence how we structure our model validation processes to ensure reliability and relevance of the predictive outcomes. [4]

An article titled, "A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games" explores the development of a machine learning model aimed at predicting NBA game outcomes using a refined team efficiency index. This index is derived from detailed statistical analysis of NBA games and incorporates elements such as basic, derived, and advanced basketball game statistics, which are augmented with league-wise data for a comprehensive feature set. The model emphasizes the importance of selecting the right features and using an optimal time window for training data to improve prediction accuracy. The concept of using an optimal time window to train models could be adapted to refine our predictive algorithms, ensuring they are trained on the most relevant and recent data, thus maintaining high accuracy and relevance across different NBA seasons. This strategy could be crucial in developing a dynamic model that adapts effectively to the fast-paced changes in team compositions and player performances in the NBA. [6]

Lastly the last study we looked at was, "*NBA Game Result Prediction Using Feature Analysis and Machine Learning*" explores the application of machine learning (ML) techniques to predict outcomes of NBA games. The research focused on identifying key statistical features from historical NBA game data that significantly influence game outcomes. The study utilized various ML methods, including Naïve Bayes, Artificial Neural Networks, and Decision Trees, to analyze and model these data points. Among the features evaluated, defensive rebounds (DRB) were identified as the most influential in determining game results. Other significant factors included three-point percentage (TPP), free throws made (FT), and total rebounds (TRB), which collectively improved the model's prediction accuracy by 2-4%. This article is directly relevant to our project, as it underscores the effectiveness of using machine learning to predict NBA game outcomes—a central aspect of our project. The focus on feature selection and analysis aligns with our project's goal to identify and utilize the most predictive elements of game data to enhance betting accuracy. By incorporating similar methodologies, such as feature analysis and testing different machine learning algorithms, our project can refine its predictive models to optimize betting strategies based on empirical data. Additionally, the findings regarding the impact of specific game features (like DRB and TPP) provide valuable insights that could guide the feature engineering process in our project, potentially leading to more accurate predictions and profitable betting outcomes.[5]

3 Methodology

Data Collection and Preprocessing

Our study utilized box score data from NBA games spanning the 2012 to 2024 seasons, obtained from basketball-reference.com. The dataset, initially in CSV format, underwent several preprocessing steps to prepare it for analysis:

- **Column Removal**: We removed columns representing minutes played ('mp' and 'mp_opp') as they were constant values (240 minutes), not contributing to model variability.
- **Season Identification**: A 'season' column was added to categorize games by their respective NBA season, facilitating time-series analysis.
- **Target Variable Creation**: A binary 'target' column was introduced to indicate whether a team won their next game, transforming our task into a binary classification problem.
- **Handling Missing Values**: Columns with null values were removed to ensure data integrity and model accuracy.
- **Feature Engineering**: We implemented rolling averages for each statistical category over the last 10 games, capturing recent team performance trends essential for our predictive models.

Model Selection and Training

We explored multiple machine learning models to predict game outcomes, selecting based on prior research and initial testing:

- **Logistic Regression**: Chosen for its baseline performance in binary classification tasks, configured with a maximum of 2000 iterations.

- **Ridge Regression**: Employed to examine regularization effects on prediction with an alpha value of 1.0.
- **XGBoost**: Configured with a binary logistic objective, a maximum depth of 3, a learning rate of 0.1, and 100 estimators, known for its robustness and efficiency in handling structured data.
- **Neural Network**: A deeper model with four layers (32, 16, 4, 1 neurons) and a dropout rate of 0.1 after the first layer, to assess performance enhancements through complexity and regularization.

All models, except for the neural network, utilized a subset of features identified as most predictive through sequential feature selection.

Training Process

The dataset was scaled using a Min-Max scaler to normalize the input features. A time series split (n=3) method was used to divide the data, preserving the time order of games, which is crucial for our time-series prediction task.

Model Evaluation

Models were primarily evaluated based on test accuracy, which directly reflects their ability to predict the outcomes of NBA games accurately. Accuracies were measured before and after rolling average features were added. This metric was selected as it provides a straightforward measure of performance in binary classification tasks.

4 Results

Our evaluation focused on the accuracy of four different machine learning models in predicting the outcomes of NBA games both before and after implementing a 10-game rolling average feature. The results are depicted in two sets of bar charts.

Initial Model Performance

Before the rolling averages were implemented, the accuracies for the logistic regression, XGBoost, and ridge regression models were as follows:

- Logistic Regression: 53.54%
- XGBoost: 53.39%
- Ridge Regression: 53.50%

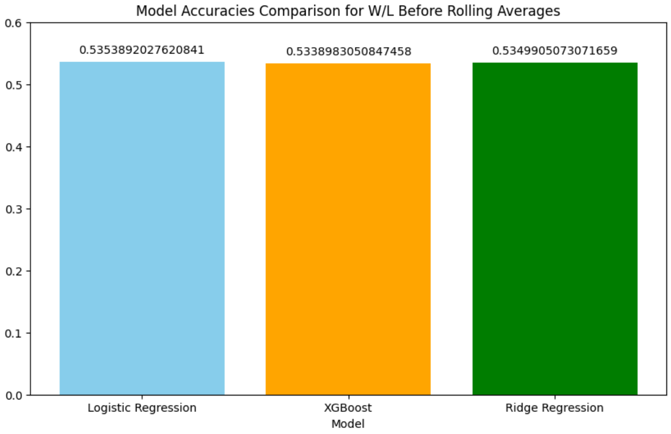

Figure 2: Initial Model Performance

These initial results were relatively close, with no model significantly outperforming the others, suggesting that without considering recent game trends, the models struggled to capture the dynamics affecting game outcomes.

Performance with Rolling Averages

After incorporating the 10-game rolling averages, a marked improvement in model performance was observed:

- Logistic Regression: Improved to 61.33%
- XGBoost: Increased slightly to 53.49%
- Ridge Regression: Improved to 61.77%
- Neural Network: Achieved an accuracy of 61.33%

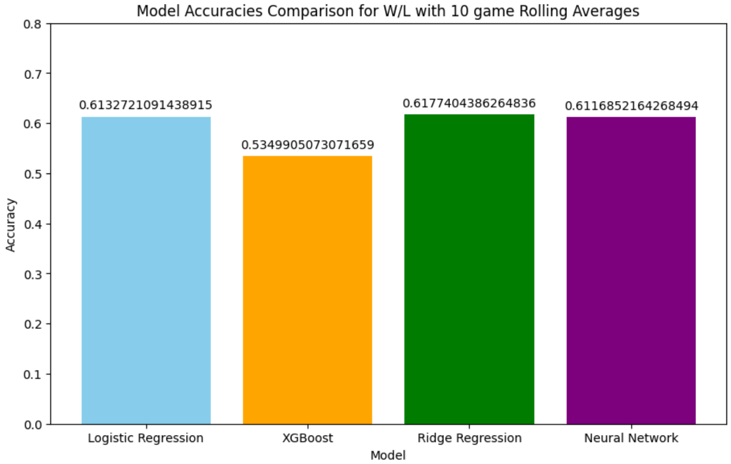

Figure 3: Model Performance with Rolling Averages

The logistic regression and ridge regression models showed the most significant improvements, suggesting that the rolling averages of team performance metrics are crucial for capturing the momentum and form of the teams, which are vital predictors of game outcomes. The neural network also benefited from the added complexity and information, matching the performance of the logistic regression.

5 Evaluation

The evaluation of the models was based solely on test accuracy because our primary objective was to assess the models' ability to correctly predict the outcomes of NBA games. Test accuracy is a straightforward metric suitable for binary classification tasks like ours.

For context, simply betting on the home team in each game would result in a 57% accuracy. This benchmark underscores the challenge of improving upon heuristic-based or simple betting strategies using machine learning models.

The dramatic improvements seen with logistic and ridge regression models, as well as the neural network, after implementing rolling averages underscore the importance of recent performance trends in predicting sports outcomes. However, even with these improved accuracies, the models achieved just over 61%, which suggests that solely relying on these models for betting might not be profitable when considering the variability and the odds set by sportsbooks.

The minimal change in XGBoost's performance might be attributed to several factors:

- **Model Complexity and Overfitting**: XGBoost is a powerful model that can capture complex patterns in data. However, without proper tuning, it might overfit less informative features before the rolling averages were introduced.
- **Feature Sensitivity**: XGBoost may require a different set of features or a different method of feature engineering to fully utilize the information in rolling averages. Its performance could also be sensitive to the hyperparameters used, such as depth and learning rate.
- **Data Dynamics**: The rolling averages, while beneficial for capturing trends, might introduce a smoothing effect that could obscure the type of rapid changes XGBoost models can capitalize on.

Additionally, while the models outperform the simple strategy of betting on the home team, the marginal gain might not suffice to offset the costs and risks associated with sports betting. Future work could explore more complex models or additional features that could potentially increase accuracy further.

Overall, the incorporation of rolling averages proved to be a decisive factor in enhancing the predictive capabilities of our models, particularly for the logistic regression, ridge regression, and neural network. More detailed tuning and exploration of feature interactions could potentially unlock higher performance in models like XGBoost.

6 Discussion

6.1 Challenges & Limitations

In developing our NBA sports betting prediction models, we encountered several challenges and limitations that impacted the overall effectiveness and accuracy of our system. These challenges highlighted potential improvement and further research:

- **Lack of Individual Player Analysis:** Our current models primarily focus on team-level statistics and do not incorporate detailed analyses of individual players' performances. This omission can lead to inaccuracies, especially in situations where key players significantly influence the game's outcome. Individual factors such as player efficiency, recent performance trends, and minute allocations are crucial for more precise predictions.
- **Ignoring Critical Factors:** Several crucial aspects of basketball games are currently not accounted for in our models:
    - **Injuries:** Player injuries dramatically affect game outcomes, and our model does not yet dynamically adjust for these changes.
    - **Trades and Roster Changes:** Player trades throughout the season can alter team dynamics significantly, yet these are

not immediately reflected in our predictive models.

  - **Team Chemistry and Coaching Impact:** The intangibles such as team chemistry and the strategic impact of coaching staff, which can vary significantly from game to game, are not considered.
  - **Home Court Advantage:** This is a well-documented factor in sports analytics that affects game outcomes, yet our model does not differentiate between home and away games in its current state.
- **Data Collection Challenges:**
  - **Tedious Scraping Processes:** Initially, gathering the necessary data was labor-intensive and time-consuming. The process of setting up reliable scraping scripts to collect data from various sources required considerable effort and maintenance.
  - **Data Cleaning and Preparation:** Transforming raw data into a usable format posed significant challenges. The data cleaning process was difficult due to inconsistencies and missing values.
- **Feature Selection Difficulties:** Identifying and selecting the most predictive features for our models was a complex and iterative process. Determining which features genuinely impacted game outcomes involved extensive testing and validation, which was both time-consuming and sometimes inconclusive.
- **Updating Data for Ongoing Seasons:** As our models rely heavily on current season data to predict outcomes accurately, the necessity to continually update the dataset with new game statistics, such as the 2023-2024 box scores, presents an operational challenge. This requirement ensures that the models reflect the most current conditions but also adds a layer of complexity in maintaining real-time accuracy.

These challenges collectively illustrate the gaps in our current modeling approach and emphasize the need for an advanced system that can dynamically integrate various data points and adjust predictions in real-time. Enhancements such as integrating player-specific analytics, improving data collection methodologies, and refining feature selection processes are crucial for overcoming these limitations. Future work will focus on addressing these challenges to enhance the robustness and accuracy of our NBA sports betting predictions.

6.2 **Future Work**

As we continue to refine and enhance our NBA sports betting project, several key areas of development have been identified to improve prediction accuracy and enrich the betting strategy. Here’s how we plan to to advance the project in the near future:

- **Incorporating More Data:** We plan to significantly expand our dataset to include comprehensive player box scores and detailed injury reports, which can drastically influence game outcomes. By capturing this data in real-time, our models will be more reflective of the current playing conditions and player availability, which are crucial for making accurate predictions.
- **Sentiment Analysis:** Leveraging sentiment analysis techniques on social media content, particularly tweets about teams and key players, will allow us to understand public sentiment and its potential impact on game outcomes. This analysis will help in adjusting the predictions based on the public perception and the psychological factors influencing the teams and players involved.
- **Sportsbook Odds Analysis:** By integrating and analyzing sportsbook odds, our project can identify value betting opportunities where there is a significant divergence between our model’s predictions and the market odds. This approach will involve developing sophisticated algorithms that can automatically compare and capitalize on these opportunities, enhancing the profitability of our betting strategies.
- **Dynamic Model Updates:** Developing a dynamic script that regularly updates our prediction models with the latest available data throughout the season is essential. This script will ensure that our models adapt to any changes in player performance, team dynamics, and other relevant factors as the season progresses, maintaining the accuracy and relevance of our predictions.
- **Playoff Game Modeling:** Adapting our models to the playoff season involves recognizing and quantifying the unique characteristics of playoff games, which often differ significantly from regular season games. Factors such as increased pressure, player experience in playoffs, and team dynamics under high-stakes conditions will be

integrated into our models to tailor predictions specifically for playoff matches.

- **Expanding Data Sources:** We will explore the incorporation of advanced metrics and external data sources such as player tracking data, which can provide deeper insights into player behaviors and game dynamics that are not captured through traditional statistics. This can enhance our understanding of the game and improve our predictive capabilities.
- **Enhanced Feature Engineering:** We will focus on refining the feature selection process to include variables that capture more subtle aspects of the games, such as player fatigue, travel schedules, and historical performance against specific opponents. This will help in building a more nuanced model that can predict outcomes with higher precision.

By addressing these areas, we intend to build a more robust, accurate and comprehensive platform for NBA sports betting prediction. These enhancements will not only improve our predictions but also offer a richer set of features and tools for bettors to utilize in their decision making process.

7 Conclusion

Throughout the development and analysis of our NBA sports betting prediction models, we have gained significant insights into the complexities of time-series prediction modeling in sports analytics. Our exploration into machine learning techniques has underscored the potential and the challenges of utilizing such technologies for betting predictions. Our study has not only highlighted the importance of rigorous data analysis and model selection but also underscored the importance for continuous adaptation and refinement of predictive algorithms to keep pace with the dynamic nature of NBA games

After using season averages and improving our data preprocessing into 10 game rolling averages we saw significant increase in accuracy in our Logistic Regression, Ridge Regression, and Neural Network models. Where we mentioned previously that our Ridge Regression model gave us the best results at predicting win losses with a prediction accuracy of 61.77% including 10 game rolling averages. We also saw the best 10 prediction features that held the most weight in our predictions which were field goals assisted, free throws ,free throws attempted, player fouls, opponent's field goal percentage, opponent's three-point percentage, opponent's offensive rebounds, opponent's steals, opponents blocks, and opponent's player fouls.

Key Discoveries:

- **Modeling Effectiveness:** Our investigation revealed that machine learning could be highly effective in sports betting, particularly for predicting NBA game outcomes. However, the accuracy and utility of these models heavily depend on the quality, breadth, and depth of the data used.
- **Feature Selection Importance:** The selection of relevant features proved crucial for the success of the predictive models. By focusing on the most influential factors like player performance metrics and team statistics, we significantly improved the prediction accuracy of our models. Features such as defensive rebounds, three-point percentages, and free throws were identified as particularly impactful.
- **Adaptability of Models:** Our project demonstrated the necessity of continuously updating and adjusting models to incorporate new game data and season-specific dynamics. This adaptability is essential for maintaining the relevance and accuracy of predictions throughout the NBA season.

In conclusion, our project has made considerable strides in applying machine learning to NBA sports betting, providing valuable insights and a strong foundation for future work in this area. By continuing to refine our approaches, expand our data sources, and explore new analytical techniques, we can further enhance the predictive capabilities of our models, contributing to the fields of sports analytics and predictive betting, and ultimately helping bettors make more informed decisions.

REFERENCES

[1] Mohammad Zubair Khan Abdullah Alsaeedi1. A study on sentiment analysis techniques of twitter data. In International Journal of Advanced Computer Science and Applications, 2019.

[2] Chenjie Cao. Sports data mining technology used in basketball outcome prediction prediction. September 2012.

[3] Filip ˇZelezn ́y Gustav Sourek, Ondrej Hubaceck. Exploiting sports-betting market using machine learning. In International Journal of Forecasting, January 2017.

[4] Fadi Thabtah Rory P. Bunker. A machine learning framework for sport result prediction. In Applied Computing and Informatics, May 2017.

[5] Zhang L. Abdelhamid N. Thabtah, F. Nba game result prediction using feature analysis and machine learning. In Annals of Data Science, January 2019.

[6] Robert Logozar ˇCaslav Livada Tomislav Horvat, Josip Job. A data-driven machine learning algorithm for predicting the outcomes of nba games. March 2023.