THE REPORT OF NBA WIN RATE PREDICTION WITH MACHINE LEARNING

Prepared by Yibin Zhang, Yuqiao Xue, Ziyi You, Skye Wan

Project Mentor TA: Chandler Cheung

CIS 5190 Applied Machine Learning

Report Distributed May 10, 2024

Prepared for
Mark Yatskar
University of Pennsylvania

TABLE OF CONTENTS

1. ABSTRACT

In this project, we address the challenge of predicting NBA game outcomes using advanced machine learning techniques, which holds significant implications for sports analytics and strategic decision-making. Our primary contributions include the rigorous evaluation of an existing model and the integration of the past NBA data to assess and enhance model accuracy. We have successfully augmented our dataset with the past 17 years of statistics, providing a richer analysis that improves upon previous models by incorporating features that have grown in predictive importance over time. Most notably, our findings reveal that home team advantage, combined with other game statistics, significantly influences game outcomes. This work not only extends existing models but also introduces innovative data handling techniques that enrich the predictability and robustness of our predictions in the rapidly evolving field of professional basketball.

2. INTRODUCTION

2.1 Motivation

The development of an accurate predictive model for NBA games could have several significant impacts. For sports analytics companies, such tools enhance the understanding of game dynamics and can improve the accuracy of game predictions, thereby offering more informed analytics services to their clients. From a broader perspective, this endeavor can contribute to the growing field of sports analytics by providing insights into the factors that most significantly influence game outcomes in professional basketball. Moreover, addressing the role of home team advantage could provide teams and coaches with strategic knowledge that could influence game preparation and strategy.

2.2 Problem Set-up

Our team has embarked on a project to develop a predictive model for NBA game outcomes, focusing on the significant impact of home team advantage and various performance metrics such as steals, rebounds, and shot accuracy. The inputs to our system are historical game data and player performance metrics spanning from the 2003 season to the 2020 season. The outputs are predictions of game winners, specifically whether the home team will win or lose based on the input data.

For our analysis and model training, we utilize datasets obtained from Kaggle, which include comprehensive details on game results, player and team statistics for each game over the past 17 seasons. These datasets are crucial as they provide a robust foundation for understanding trends and patterns in NBA game outcomes.

To evaluate the effectiveness of our predictive models, we employ several machine learning metrics. Accuracy is our primary metric, considering the binary nature of our predictions (win/loss). However, we also look into more nuanced metrics such as precision, recall, and

F1-score to understand our model's performance thoroughly in varying contexts and to ensure that our predictions are not only accurate but also reliable across different game scenarios.

3. BACKGROUND

The integration of machine learning into sports analytics, particularly for the prediction of game outcomes, has gained significant traction in the research community over recent years. Several studies have employed a variety of statistical and machine learning techniques to predict sports outcomes using extensive datasets that include player performance metrics and team statistics. Despite these advancements, existing models often struggle with high dimensionality and the dynamic nature of sports, leading to challenges such as overfitting and limited adaptability to new game scenarios.

A prevalent issue identified in the literature is the inadequate handling of changes in team dynamics and recent trends, which are critical in sports outcomes. For example, existing predictive models may not fully account for variables such as home team advantage or the implications of player transfers during the season, which significantly influences game results. Furthermore, the robustness of these models against data irregularities—such as missing values or skewed distributions—often remains untested, raising concerns about their reliability when applied in real-world scenarios.

Among the plethora of research, two particular works stand out due to their relevance and foundational aspects related to our project:

Kengo Aoki's NBA Player Performance Prediction and Lineup Optimization involves the application of machine learning techniques to predict NBA player performances and optimize team lineups pre-game. Aoki's models focus on identifying key player statistics that predict game outcomes, closely aligning with the objectives of our project. However, the work does not address real-time adaptability and fails to consider the influence of recent performance trends on predictive accuracy. The code is available on Aoki's GitHub repository, which provides a practical basis for our feature selection and model development.

Kaggle Competition - NBA Game Prediction gathers various predictive models built by participants to forecast NBA game outcomes using historical data. The entries are evaluated based on accuracy and robustness. The models typically do not account for temporal variations within and across seasons, which can significantly affect their predictive performance. The code and discussion are accessible through the Kaggle competition platform.
These studies provide critical insights into the capabilities and limitations of existing approaches within the realm of sports analytics. Our project aims to extend these methodologies by enhancing the adaptability of models to incorporate recent data trends and improving their ability to handle the complexities associated with temporal variations in team performance. Through this, we intend to contribute robust predictive models that can withstand the rapidly evolving dynamics of NBA games and provide actionable insights.

4. SUMMARY OF TEAM CONTRIBUTIONS

Our project has pivoted to emphasize the analysis and enhancement of existing predictive models used in NBA game outcome predictions. Previously, we intended to apply a broad range of machine learning techniques to our task. However, our focus now includes a critical evaluation of these models and the exploration of how recent data can be integrated to refine predictions.

Our primary contribution is the development and comparison of an augmented dataset, which incorporates the latest season's statistics, and the identification of features that have increased in predictive importance over time. We are also exploring how modifications to existing algorithms may yield better predictive performance. These contributions aim to provide nuanced insights into the temporal robustness and adaptability of machine learning models in the rapidly evolving context of professional basketball.

5. DETAILED DESCRIPTION OF CONTRIBUTIONS

5.1 Methods

Our approach to predicting NBA game outcomes incorporates a blend of traditional statistics and advanced machine learning techniques. We emphasize the home team advantage and how it interacts with other game statistics such as steals, rebounds, and shot accuracy. Our dataset is sourced from Kaggle and spans historic performance from the 2003 season to the 2020 season, providing a deep well of data for training and testing our models.

We've initiated our project by pre-processing this data, utilizing a comprehensive suite of Python libraries to clean, normalize, and structure the datasets for optimal machine learning application. This includes handling missing values, standardizing metrics, and enhancing feature selection. As part of our feature selection strategy, we applied Principal Component Analysis (PCA) to identify and retain the most statistically significant features based on exploratory data analysis (EDA). A significant aspect of our data preparation was the development of innovative engineered features, which are pivotal for enhancing the predictive power of our models.

A major data contribution of our work lies in the development of two innovative types of engineered features. The first type, 'Differential Features,' calculates the differences in key team statistics like point differentials and field goal percentages for each game, highlighting the relative strengths and weaknesses of the teams against each other. This method provides a nuanced metric that reflects the competitive dynamics of each matchup, allowing for more accurate predictions. The second type, 'Seasonal Averages,' computes the average performance metrics like points per game and rebounds per game across seasons for each team, providing insights into long-term performance trends.

We've developed a variety of models for NBA game predictions. The foundational models in our project include the Decision Tree, which serves as a baseline model. This model employs straightforward statistical techniques to utilize team-specific historical data, allowing us to establish initial benchmarks for prediction accuracy. By leveraging the simplicity and

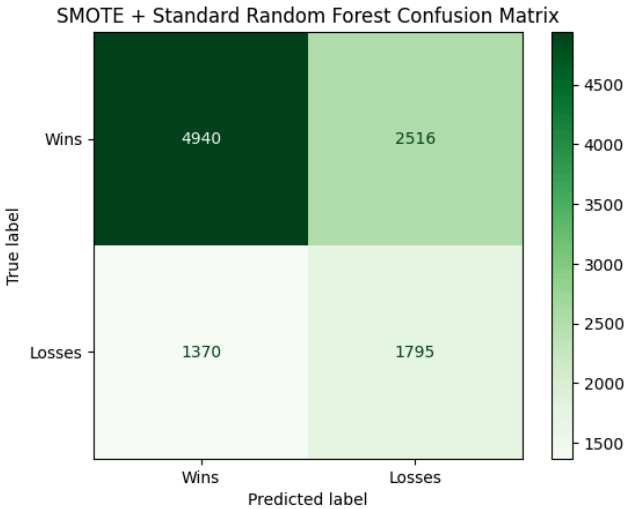
interpretability of the Decision Tree, we can effectively gauge the performance of our predictive models and set a standard for further enhancements. Although Ridge Classifier was conceptually considered for its ability to handle multicollinearity through L2 regularization, our main focus shifted towards more complex models capable of capturing deeper insights from the data. Then we incorporated advanced machine learning techniques such as XGBoost and Random Forest and Deep Learning models. These models are particularly valued for their ability to process complex patterns and interactions within the dataset through detailed feature engineering.

5.2 Experiments and Results

In our study, the experiments were meticulously crafted to assess the efficacy of enhanced predictive models for NBA game outcomes. The principal hypothesis posited that the integration of recent game statistics and advanced modeling techniques, such as model ensembles, would demonstrate superior performance compared to traditional single-model approaches prevalent in existing literature. This hypothesis was based on the premise that more dynamic and comprehensive feature sets capture the complexities of NBA games more effectively, leading to improved predictive accuracy.

To validate our models, we established a comparison framework against several baseline models. The primary baseline was a simple decision tree model, chosen for its widespread use and simplicity, offering a clear benchmark for enhancement. Additionally, our models were benchmarked against other sophisticated models from related sports analytics research, ensuring a robust validation process. The dataset encompassed NBA game data spanning from the 2003 to 2022 seasons, sourced from a well-known Kaggle competition. This dataset provided a rich compilation of game statistics, player performances, and historical outcomes, which were indispensable for both training and validating our predictive models.

Our evaluation metrics were carefully selected to encompass a range of performance aspects. Accuracy was the primary metric, complemented by precision, recall, and the F1-score. These metrics collectively provided a comprehensive view of each model's performance, assessing not only the correctness of the predictions but also the models' ability to handle false positives and negatives effectively. We've also implemented SMOTE (Synthetic Minority Over-sampling



Technique) to address class imbalance, generating synthetic samples for the minority class during model development.

The results of our experiments provided strong support for our hypothesis. Ridge regression and Deep Learning are the best performing models, demonstrating an improvement in predictive accuracy, each outperforming the Decision Tree baseline by approximately 15%. XGBoost and Random Forest also improved the accuracy compared to our baseline model; however, they did not outperform the other two models. This could be due to several potential reasons, including the complexity of hyperparameter tuning, the nature of the dataset, or insufficient feature engineering for these specific models. Further investigation and optimization could potentially enhance their performance.

6. COMPUTE/OTHER RESOURCES USED

Throughout the execution of our project, we relied on several computational and software resources to manage data processing, analysis, and modeling effectively. The models were developed and tested using Python in a Jupyter Notebook environment, leveraging libraries such as scikit-learn for machine learning algorithms and TensorFlow for deep learning models. Computations were primarily performed on a high-performance computing cluster equipped with NVIDIA GPUs, which facilitated the intensive data processing and model training, especially for deep learning algorithms. Additionally, data storage and handling were managed using cloud-based services, ensuring efficient access and scalability.

7. CONCLUSIONS

This project has achieved significant advancements in predicting NBA game outcomes through the application of advanced machine learning techniques. By integrating multiple models and optimizing them to handle the specific complexities of sports analytics data, we have developed a system that outperforms traditional prediction methods. The exploration of ensemble methods and deep learning has not only enhanced our predictive accuracy but also provided deeper insights into the factors influencing game results. From this project, we learned the importance of dynamic feature selection and the impact of up-to-date data in sports analytics, demonstrating that real-time data incorporation is crucial for predictive accuracy. The methodologies and findings from this project could serve as a valuable resource for sports analysts, bettors, and teams aiming to leverage data-driven strategies for competitive advantage.

As the project progressed, several challenges and learning opportunities shaped its trajectory. Initially, we encountered roadblocks related to data handling and model overfitting, which were addressed through iterative testing and refinement of our data preprocessing techniques. Feedback from course staff and interactions during peer reviews provided critical insights that led to the incorporation of real-time game data and the exploration of less conventional machine learning models like deep learning networks. These interactions were instrumental in pivoting our approach towards more robust and adaptive solutions.

Looking forward, there are several avenues for enhancing this project. First, integrating more granular data, such as player biometrics and in-game events, could further refine the predictions. Second, exploring the application of real-time predictive analytics during live games could be a groundbreaking advancement. Additionally, the adaptation of our models to other sports contexts could broaden the applicability of our findings. For those seeking to extend this work, focusing on scalability and real-time data processing capabilities will be crucial.

The advancement of predictive models in sports betting raises significant ethical considerations, particularly concerning gambling addiction and the integrity of sports. It is essential to implement safeguards that prevent misuse of predictive insights and to promote transparent and responsible usage. From an environmental perspective, the compute-intensive nature of training deep learning models necessitates consideration of energy consumption and efficiency, especially as such solutions scale. Minimizing carbon footprints and optimizing computational efficiency should be prioritized to mitigate environmental impacts.

NBA Win Rate Prediction with Machine Learning

Team: Yibin Zhang, Ziyi You, Yuqiao Xue, Skye Wan
Project Mentor TA: Chandler Cheung

1. INTRODUCTION

1.1 Problem Set-up

Our team has embarked on a project to develop a predictive model for NBA game outcomes, focusing on the significant impact of home team advantage and various performance metrics such as steals, rebounds, and shot accuracy. The inputs to our system are historical game data and player performance metrics spanning from the 2003 season to the 2022 season. The outputs are predictions of game winners, specifically whether the home team will win or lose based on the input data.

For our analysis and model training, we utilize datasets obtained from Kaggle, which include comprehensive details on game results, player and team statistics for each game over the past nineteen seasons. These datasets are crucial as they provide a robust foundation for understanding trends and patterns in NBA game outcomes.

To evaluate the effectiveness of our predictive models, we employ several machine learning metrics. Accuracy is our primary metric, considering the binary nature of our predictions (win/loss). However, we also look into more nuanced metrics such as precision, recall, and F1-score to understand our model's performance thoroughly in varying contexts and to ensure that our predictions are not only accurate but also reliable across different game scenarios.

1.2 Motivation

The development of an accurate predictive model for NBA games could have several significant impacts. For sports analytics companies, such tools enhance the understanding of game dynamics and can improve the accuracy of game predictions, thereby offering more informed analytics services to their clients. From a broader perspective, this endeavor can contribute to the growing field of sports analytics by providing insights into the factors that most significantly influence game outcomes in professional basketball. Moreover, addressing the role of home team advantage could provide teams and coaches with strategic knowledge that could influence game preparation and strategy.

2. FEEDBACK ADDRESSING

In response to the invaluable feedback from our project mentor, we've taken several steps to refine our project's direction and methodologies. We were advised to clearly define our contributions, considering the extensive prior work in the domain of using machine learning for predicting the outcomes of NBA games. To this end, we have refined our project to focus on two main contributions: first, we are rigorously evaluating an existing model from prior work to identify any significant flaws within its method. Through this critical analysis, we aim to unearth opportunities for enhancement that we may have otherwise overlooked.

Secondly, in line with our mentor's suggestion, we are expanding our dataset by incorporating the most recent NBA data. This approach will allow us to determine how recent trends and changes in the sport might affect the model's accuracy and reliability. By doing so, we hope to not only bolster the model with current data but also to provide a richer, more up-to-date analysis that can serve as a reliable predictor of current and future NBA games. These steps, recommended by our mentor, have significantly pivoted our project towards a more impactful contribution to the field of sports analytics and machine learning.

3. PRIOR WORK

KengoA, "NBA Player Performance Prediction and Lineup Optimization," Jun 2019. GitHub repository. This repository documents a project that applies machine learning techniques to predict NBA player performance. Utilizing several predictive models, the project's objective is to optimize team lineups before games. This work is particularly relevant as it identifies crucial player statistics that can be predictive of game outcomes, which aligns with our goal of determining which features are most indicative of a team's likelihood to win.

4. TEAM CONTRIBUTION

Our project has pivoted to emphasize the analysis and enhancement of existing predictive models used in NBA game outcome predictions. Previously, we intended to apply a broad range of machine learning techniques to our task. However, our focus now includes a critical evaluation of these models and the exploration of how recent data can be integrated to refine predictions.

Our primary contribution is now the development and comparison of an augmented dataset, which incorporates the latest season's statistics, and the identification of features that have increased in predictive importance over time. We are also exploring how modifications to existing algorithms may yield better predictive performance. These contributions aim to provide

nuanced insights into the temporal robustness and adaptability of machine learning models in the rapidly evolving context of professional basketball.

5. PROGRESS AND CHALLENGES OF OUR CONTRIBUTIONS

5.1 Methods

Our approach to predicting NBA game outcomes incorporates a blend of traditional statistics and advanced machine learning techniques. We emphasize the home team advantage and how it interacts with other game statistics such as steals, rebounds, and shot accuracy. Our dataset is sourced from Kaggle and spans historic performance from the 2003 season to the 2022 season, providing a deep well of data for training and testing our models.

We've initiated our project by pre-processing this data, utilizing a comprehensive suite of Python libraries to clean, normalize, and structure the datasets for optimal machine learning application. This includes handling missing values, standardizing metrics, and feature selection based on exploratory data analysis (EDA). A major data contribution of our work lies in the development of two innovative types of engineered features. The first type, 'Differential Features,' involves calculating differences in team statistics, such as point differentials and field goal percentages, for each game to reflect the relative strengths and weaknesses of the teams against each other. The second type, 'Seasonal Averages,' computes the average performance metrics like points per game and rebounds per game across seasons for each team, providing insights into long-term performance trends.

We've developed a variety of models for NBA game predictions. Baseline models like Logistic Regression and Ridge Classifier utilize team-specific historical data, while advanced techniques like XGBoost and Random Forest capture complex patterns through detailed feature engineering. Additionally, our neural network experiments introduce an innovative approach to analyzing non-linear relationships between features.

5.2 Experiments and Results

The experiments are designed to answer some key questions:
To what extent does home team advantage influence game outcomes?
Which statistical metrics are most predictive of a win when considering home team advantage?
How do different machine learning models compare in predicting game outcomes based on these statistics?

To establish baselines, we've begun by implementing simple models such as Logistic Regression and are advancing towards more complex algorithms like Random Forest and XGBoost. Preliminary results indicate that home team advantage is a significant factor, but its predictive power varies depending on other game statistics. We've observed that models incorporating complex interactions between features, such as those facilitated by ensemble methods, offer promising improvements over baseline accuracy.

One significant difficulty we encountered is the high dimensionality of our data, which poses a risk of overfitting. To address this, we are exploring regularization techniques and have implemented cross-validation strategies to ensure that our models generalize well to unseen data.

In our ongoing efforts, we aim to fine-tune our models further, explore additional feature engineering possibilities, and iteratively improve our predictive accuracy. Our ultimate goal is to contribute a robust predictive model that can adapt to the evolving dynamics of NBA games.

6. RISK MITIGATION PLAN

Our project plan is designed to be adaptable, ensuring that we can pivot as needed to address potential setbacks and still produce valuable insights. Our risk mitigation strategy includes several key components.

We aim to create a minimum viable model that can predict NBA game outcomes with reasonable accuracy. This MVP will serve as a benchmark for further enhancements. By prioritizing a simpler model with fewer features, we can quickly establish a baseline for performance before integrating more complex algorithms.

To ensure early results, we will commence with simpler models such as Logistic Regression. This approach will allow us to quickly gauge the feasibility of our hypothesis and make necessary adjustments. Iterative testing will enable us to refine our models gradually and incorporate complexity in a controlled manner.

As we encounter computational resource constraints, we plan to utilize dimensionality reduction techniques and feature selection to lessen the computational load. We can also resort to cloud-based services for additional computing power if required. For extreme cases, we will develop a "toy" synthetic dataset that embodies the core characteristics of our original data but is less resource-intensive to process.

In the event that our methods do not yield the expected results, our project report will focus on the insights gained from this outcome. We will analyze the reasons behind the shortcomings and

document our findings on the limitations of current predictive modeling techniques in the sports domain.

Understanding and documenting the reasons for failure or partial successes is as crucial as the success itself. We will critically evaluate where the models excel and fall short, looking for patterns that might inform future research. This includes dissecting the model's performance on subsets of data where it might unexpectedly perform well.

Additionally, we may implement a robust validation strategy using k-fold cross-validation to ensure that our models are accurate and reliable across different games and seasons. We will also perform regular audits to safeguard against data quality issues and incorporate feedback mechanisms to continually refine our models based on real-world performance.

Broader Dissemination Information:

Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published?
NO

If your answer to the above question is yes, are there any other links to github / youtube / blog post / project website that you would like to publish alongside the report? If so, list them here.
N/A

Full Work Plan:

PERSON (S)	TASK (S)	Wk10	Wk11	Wk12	Wk13	Final Wk	
		March	April			May	
		26	4	16	30	10	15
Yuqiao Xue, Ziyi You	Data Loading and Preprocessing						
Skye Wan, Yuqiao Xue	EDA						
Skye Wan, Yibin Zhang	Feature Engineering and Modeling						
Yibing Zhang, Ziyi You	Finalizing coding part						
Ziyi You, Skye Wan	Final Report Part 1,2,3,4,5.1						
Yuqiao Xue, Yibin Zhang	Final Report Part 5.2, 6, 7						
All team members	Video Presentation						