Stats 101C Final Project

Predicting NBA Game Results

– – Group 24

Yangsheng Xu, Cecilia Liu, Sizhuo Tian, Charles Wei (806282612), Chengrui Hao

Table of Contents

1 Introduction

1.1 Background of the Problem

The **2023 NBA regular season** saw fierce competition among **30 teams**, culminating in a dataset of **2,460 games**. Each game is characterized by **24 features**, capturing key statistics such as points scored, field goals, assists, and rebounds.

Meanwhile, predicting basketball game outcomes has always been a hot topic among fans. Most people rely on **intuition** to classify teams into strong and weak categories, they firmly believe that strong teams will defeat weaker ones. While a few are deeply engrossed in **data**, using numbers to determine the winners.

Based on this dataset, we aim to consider **six factors** - **Average Score Difference, Weighted Average of a Team's Historical Points Scored, Home Advantage, Consecutive Wins/Losses, to be filled, and Previous Head-to-head Results** to build a predictive model, striving for higher prediction accuracy. Also, understanding the dynamics that lead to winning or losing a game helps teams make informed decisions regarding strategy and player performance improvements.

1.2 Restatement of the Problem

- Split the dataset into training data and test data, the cutoff time period is Feb 26, 2023 (Games before and on the cutoff date will be the training data, and afterward will be test data)
- Breakdown the model into training 6 key factors
- Combine the factors into a full model
- Predict game results after Feb 26 and calculate accuracy using real data

1.3 Glossary/Define Key Terms and Variables

- Head-to-head Result/Record: Previous competition results between two teams
- W/L: Win or Loss indicator (W or L)
- MIN: Minutes played
- PTS: Points scored
- FGM: Field goals made
- FGA: Field goals attempted

- FG%: Field goal percentage
- 3PM: Three-point field goals made
- 3PA: Three-point field goals attempted
- 3P%: Three-point field goal percentage
- FTM: Free throws made
- FTA: Free throws attempted
- FT%: Free throw percentage (some entries are non-numeric)
- OREB: Offensive rebounds
- DREB: Defensive rebounds
- REB: Total rebounds
- AST: Assists
- STL: Steals
- BLK: Blocks
- TOV: Turnovers
- PF: Personal fouls
- +/-: Plus/minus statistic

2 Identify and Justify Assumptions

- Player transfers have no significant impact on the team's subsequent win rate
- All teams aim to win the game, and there is no situation of intentionally underperforming
- The performance of a team is stable, and can be fully represented by statistics

3. Data Processing

In the data processing stage, we began by checking for any missing values (NA) in the dataset and found none. The only processing step required was converting the date column into the appropriate Date format in R to ensure that it was recognized correctly for further analysis. No additional cleaning or transformations were necessary.

In the data processing stage, several key features were added to the dataset to enhance the model's predictive capability. These included `Avg_Score_Diff`, which quantifies the scoring difference between a team and its opponent; `Weighted_Avg_PTS`, emphasizing recent performance trends by assigning higher weights to recent games; `HomeWinRate`, capturing the advantage of playing at home; and streak-based features like `Consecutive_Win_General` and `Consecutive_Loss_General` to account for team momentum. Additionally, comparative metrics such as `Point_Differential`, `Shooting_Differential`, and `Turnover_Differential` were included to highlight specific aspects of team strategy and execution. These features were derived from the existing dataset and tailored to quantify game dynamics, forming the foundation for the model's analysis and predictions.
We also separate the dataset into training and testing data set with an edge of date 2.26.2024. In this way, the data set is splitted into 70% and 30% and also the length of the training dataset

include the deadlinel of trading, which include our model’s accuracy since some big trade is very likely to happen during this period and containing them will help improve our accuracy.

4 Modeling

To increase the model's accuracy, our group attempted to mine the hidden information in the dataset.. We breaks the full model into 6 parts：

4.1 Average Score Difference (Avg_Score_Diff)

Avg_Score_Diff is calculated as the difference between the average points scored by the team and the average points scored by their opponent.This feature is chosen because it directly quantifies the relative scoring strength of one team compared to their opponent, making it highly relevant to predicting game outcomes.

The goal of this algorithm is to predict the winning team based on the **average total scores**. Specifically, we hypothesize that the team with a higher average total score is more likely to win.

The algorithm first calculates the historical average points for each team based on their past games. This step captures the overall scoring ability of each team. Next, it identifies the opponent for each game and fetches their historical average points scored. Using this information, the algorithm computes the score difference for each game by subtracting the opponent's average score from the team's average score.
Once the features are ready, the dataset is split into training and testing sets based on a specific cutoff date(before 2024/2/27). This ensures that the model is evaluated on unseen games to simulate real-world predictions. A Random Forest classifier is then trained using the Avg_Score_Diff feature to predict game outcomes, specifically whether a team wins or loses.

The model achieved an accuracy of 58.24%, which indicates moderate performance in predicting game outcomes based on the average score difference. While the predictions capture some patterns in the data, the misclassification rates suggest that additional features or alternative methods may be needed to improve predictive power.

4.2 Weighted average of a team's historical points scored

The feature used for this task is the weighted average of a team's historical points scored, with more weight assigned to games closer to the current game date. This feature is designed to capture the team's recent performance trends, as more recent games are likely to reflect the team's current form, strategies, and player conditions more accurately than older games. By assigning higher weights to more recent data, the model emphasizes these recent trends while still considering historical performance.

The algorithm calculates this feature by first computing the weights for each past game as the inverse of the days between the past game and the current game date. These weights are then used to compute a weighted average of the team's points scored. If no past games exist for a team, the team's average score across all games is used as a default value. This ensures that the feature remains informative even for teams with limited historical data.

The model is trained using the weighted average as the predictor for the game's outcome (win or loss). A Random Forest classifier is applied to learn the relationship between this feature and the target variable. The model is evaluated using a test set split by a specific cutoff date(before 2024/2/27) to simulate future game predictions.
The results show an accuracy of 52.84%, which is relatively low. This suggests that while the weighted average of recent points captures some aspects of team performance, it may not fully explain game outcomes. Factors such as opponent strength, home-court advantage, or specific game-day conditions might play a significant role and could be considered to improve the model's accuracy.

4.3 Home Advantage

It is expected that NBA teams always tend to win at their home game. The famous home-advantage teams are the boston celtics and los angeles lakers. These two teams significantly increase their winning chance while they are playing at home. Hence, it is reasonable to include such a factor in our model.
To achieve this, we create a new column for our training data(Before date 2.26), which calculates the average home win rate of each NBA team . For teams with no prior home game data, we assigned a default win rate (e.g., 0.5) to ensure completeness for testing purposes.
A Random Forest model was then trained on the processed data, with Higher Win Rate (a binary indicator of whether the home win rate exceeded 0.5) as the main predictor and the game outcome (W/L) as the target variable. The testing dataset (games on or after February 26, 2024) was used for validation. The model achieved a prediction accuracy of **62.03%** with a balanced accuracy of 61.52%. The confusion matrix revealed 77 true negatives, 155 true positives, 37 false negatives, and 105 false positives. While the model's specificity was strong (80.73%), sensitivity was moderate (42.31%), indicating the model performed better at predicting wins than losses. Despite the limitations, this simple home-field advantage feature demonstrates predictive value and highlights the effectiveness of Random Forest in leveraging domain-specific metrics.

4.4 Consecutive Wins/Losses

We expected that if a team had won consecutively prior to the predicted game, it would increase the team’s chances of winning the next game, and the same logic would apply to consecutive losses decreasing the chances of winning. To explore whether consecutive wins or losses could influence a team's likelihood of winning or losing the next game, we engineered four new features in the dataset: Consecutive_Win_Between, which counts consecutive wins against the same opponent in the exact order of matchups; Consecutive_Loss_Between, which tracks consecutive losses under the same conditions; Consecutive_Win_General, representing a team's general consecutive win streak prior to a game; and Consecutive_Loss_General, capturing a team's general consecutive loss streak prior to a game. However, due to the limited number of cases where teams achieved consecutive wins or losses, these features did not prove useful for predicting game outcomes on itself. When running a logistic regression model on these four added columns, none of these features were statistically significant, with p-values close to 1, indicating no relationship with the target variable. We also trained a Random Forest

model using these features, but the accuracy was only 50%, which is equivalent to random guessing. This suggests that while consecutive streaks may intuitively seem important, they were not predictive in this context due to the sparsity of such occurrences in the dataset.

4.5 Offensive/Defensive Efficiency

To capture a team's overall performance trends and integrate meaningful game metrics, we engineered three key features: Point Differential, Shooting Differential, and Turnover Differential. These metrics were calculated by comparing a team's game performance against its opponent in terms of scoring, shooting efficiency, and turnovers. For instance, the Point Differential quantifies the scoring margin in each game, which directly reflects a team's dominance or struggle, while the Shooting Differential highlights disparities in shooting accuracy between the team and its opponent. Finally, the Turnover Differential sheds light on ball-handling effectiveness relative to the opposition. We then extended these metrics to represent historical team trends by calculating their average values for each team. These averages were added to the dataset to ensure pre-game information availability and enhance feature consistency, helping the model focus on broader performance patterns rather than individual game fluctuations.

4.6 Previous Head-to-head results

Analyzing **previous head-to-head results** helps predict game outcomes by revealing patterns in team dynamics, strengths, and weaknesses. Past encounters highlight matchup-specific factors, such as consistent dominance, strategic advantages, or key player performances. Intuitively, if one team loses to the other team frequently, one would guess that there is some **kind of restraint** between them and continue to guess future winners and losers based on past win/loss results; conversely, if there is a consistent pattern of losses and wins between the two teams, the predictive value of the feature decreases.

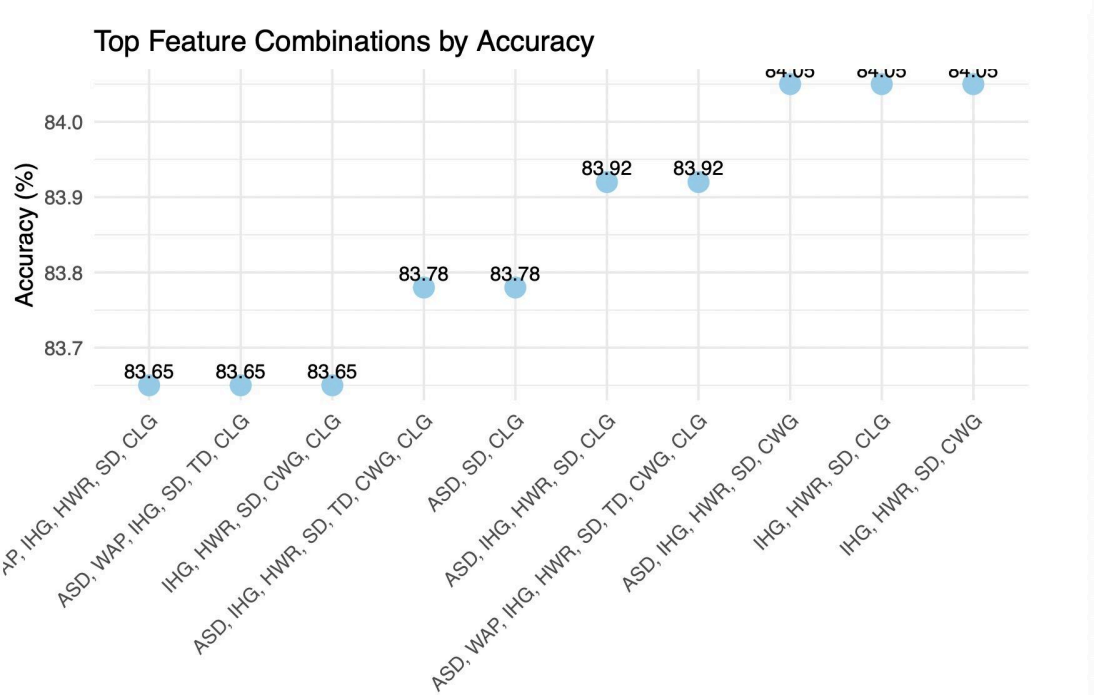
We first tried to train a **linear regression model** using previous head-to-head results. To use linear regression for predicting game results, historical data including **game outcomes and key statistics like points scored, rebounds, assists, and turnovers** are selected from the whole dataset. The target variable (e.g., points differential or win/loss as binary) is defined, and the features serve as explanatory variables. The model learns a linear relationship between the features and the target, estimating coefficients that indicate each statistic's influence. The accuracy using this linear regression model solely results in **an accuracy of 85%**.

We also tried to use an **SVM model** to analyze previous head-to-head results. Key performance metrics include **PTS, FGM, FGA, 3PM, OREB, DREB, AST, STL, BLK, TOV, PF** from the dataset. The target variable was win/loss (binary), while the features represented game statistics. Using SVM with a linear kernel, the model identified an optimal hyperplane that separates winning and losing outcomes based on the feature set. Training data included games played before Feb 26, allowing the model to learn patterns and relationships. The trained SVM was then tested on future games to evaluate its accuracy, resulting in a **89.05% accuracy**, which is higher than using linear regression.

4.7 Final Model

The Random Forest model was trained using all possible combinations of these features, and the top-performing combination was identified based on testing accuracy. Results showed that including features like `Avg_Score_Diff`, `HomeWinRate`, and `Shooting_Differential` in the model consistently improved accuracy, with the best feature set achieving an accuracy of 84.05%. This highlights the importance of combining scoring trends, home-court advantage, and game execution metrics for reliable predictions. The model demonstrates strong predictive power while remaining interpretable and robust, suitable for capturing key factors influencing NBA game outcomes.

Then we visualize the result with graph which help increase the efficiency of our report (We have attached an abbreviation table for newly created column at the reference list):



5 Interesting Findings:

One of the most notable findings in this project was the impact of `HomeWinRate` on predicting game outcomes. Teams with higher home win rates consistently demonstrated an advantage, reinforcing the well-documented "home-court advantage" phenomenon in sports. Another key observation was the moderate predictive power of `Avg_Score_Diff`, which highlights scoring efficiency as a crucial factor in determining game outcomes. However, features like `Consecutive_Win_General` and `Consecutive_Loss_General` had limited predictive utility due to the sparsity of streak occurrences in the dataset, suggesting that momentum may

not play as significant a role as hypothesized. However, when combined with other factor, the outcome accuracy slightly improves. Lastly, the `Shooting_Differential` and `Turnover_Differential` metrics provided valuable insights into a team's execution and ball-handling, which strongly correlated with winning outcomes.

6 Strength and Weaknesses of Model

A notable strength of using a random forest model to predict NBA game results is its ability to handle complex, nonlinear relationships and interactions between variables such as prior game results, scores, home advantage, and team performance. By creating an ensemble of decision trees, each trained on a random subset of the data and features, the model effectively reduces the risk of overfitting and provides robust predictions even with noisy or missing data. For instance, the model can capture nuanced patterns, such as how a team's performance at home differs significantly from their performance on the road or how the margin of victory in previous games correlates with future outcomes. Additionally, the random forest's feature importance metrics enable insights into which factors—such as recent scoring trends or historical performance against specific opponents—most strongly influence the prediction. This interpretability, combined with the model's high adaptability to diverse datasets, makes it a powerful tool for predicting game results with high accuracy and reliability.

However, a significant weakness of this model is that it relies exclusively on prior game statistics, disregarding crucial game-specific factors that can greatly influence outcomes. For example, the model does not account for injuries, which can drastically alter a team's performance. If a star player or a critical contributor is sidelined, the team's chances of winning may decrease, and the absence of this information in the model could lead to inaccurate predictions. Additionally, the model overlooks strategic elements such as a team's decision to rest players for future matches or modify their tactics based on the opposition. These strategies are often influenced by a team's position in the standings; for instance, a team fighting for a playoff spot may be highly motivated and play at peak intensity, whereas a team with a secure position might prioritize player health and experiment with lineups. Without incorporating such dynamic, context-specific variables, the model risks over-relying on historical patterns that may not reflect the unique circumstances of individual games, ultimately limiting its predictive accuracy.

7 Conclusion

In conclusion, the final Random Forest model effectively combined features such as `Avg_Score_Diff`, `HomeWinRate`, and execution-based metrics to predict NBA game outcomes with an accuracy of 84.05%. This performance underscores the value of integrating statistical trends, contextual factors like home-court advantage, and dynamic performance indicators. While the model captures many key aspects of team performance, it is limited by its reliance on historical data, excluding real-time factors like injuries or tactical changes. Future iterations could improve predictive accuracy by incorporating game-day specifics and

player-level metrics. Overall, the model serves as a robust starting point for predictive analysis in sports.

8 Reference List

Table 1: Abbreviation

Abbreviation	Meaning
ASD	Avg_Score_Diff
WAP	Weighted_Avg_PTS
IHG	IsHomeGame
SD	Shooting_Differential
TD	Turnover_Differential
CWG	Consecutive_Win_General
CLG	Consecutive_Loss_General
HWR	HomeWinRate