STATS 101C Final Project

Yangsheng Xu, Cecilia Liu, Sizhuo Tian, Charles Wei, 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** that capture key statistics such as points scored, field goals, assists, and rebounds.

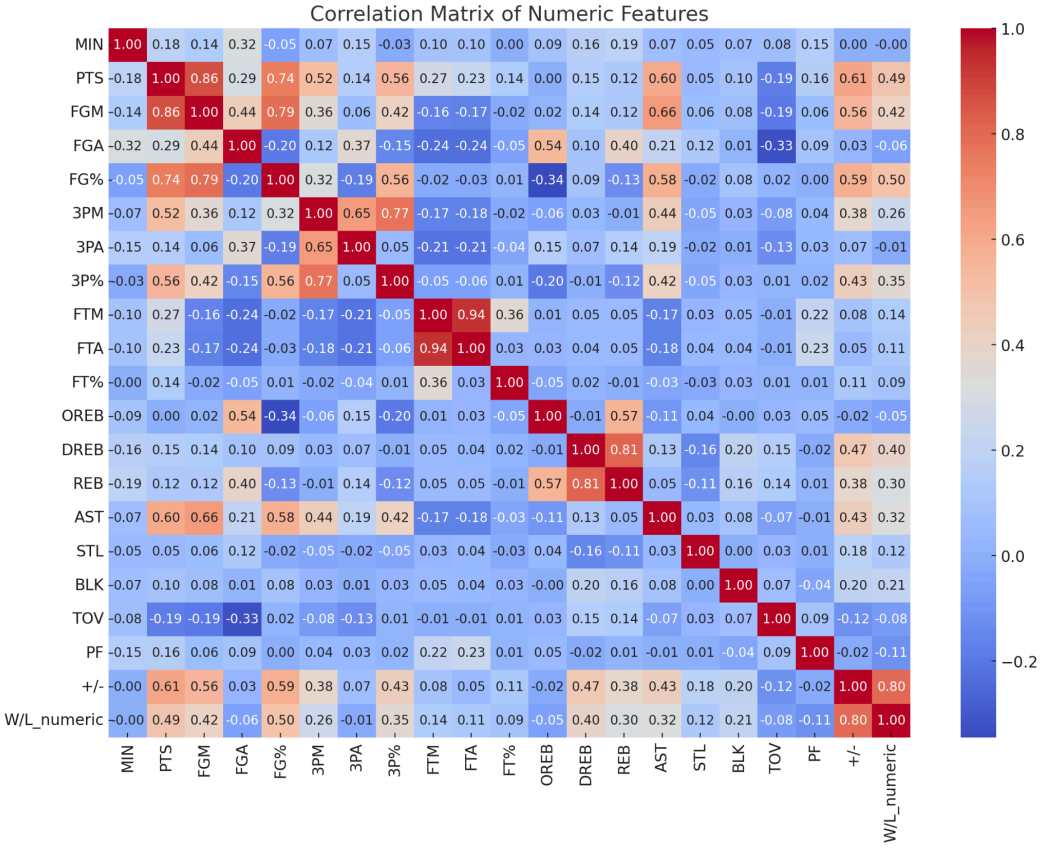
Meanwhile, predicting basketball game outcomes has always been a hot topic among fans. Most people rely on **intuition** to classify teams into strong and weak categories, believing that strong teams will defeat weaker ones. While a few are deeply engrossed in **data**, using numbers to determine the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

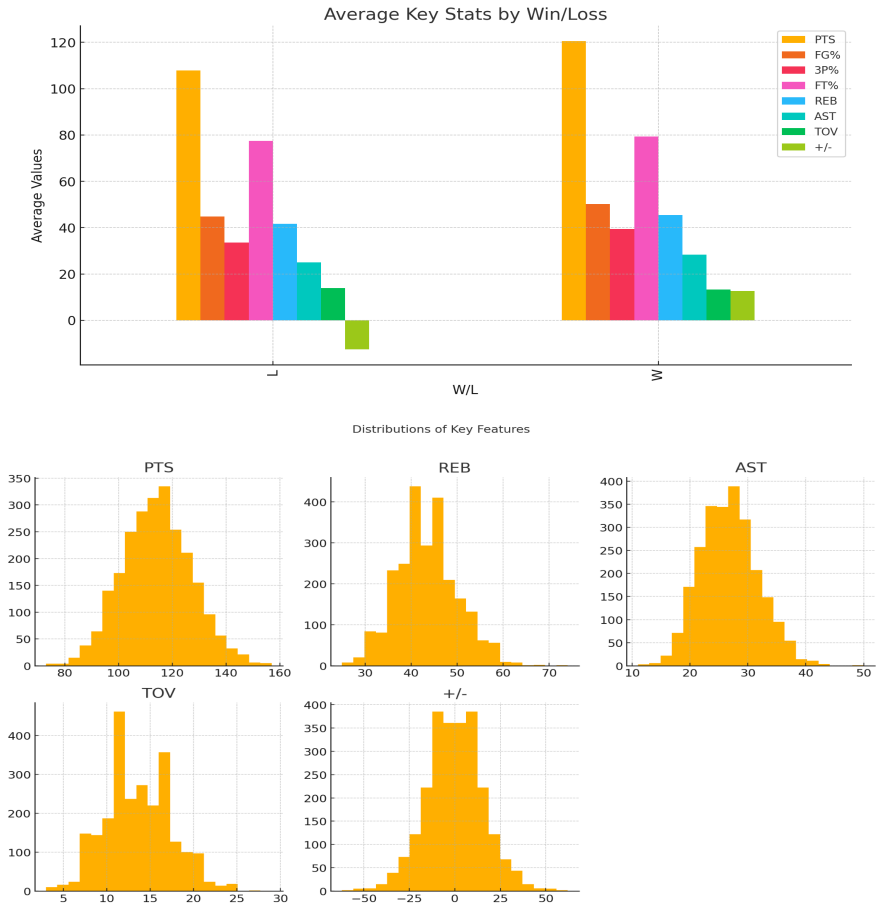
1.2 Restatement of the Problem

The task involves using historical game data to predict NBA game outcomes. A cutoff date of February 26, 2023, is selected to divide the dataset into two parts: all games played on or before this date form the training set, while all subsequent games form the test set. The approach first identifies six key predictive factors from the training data. These factors will then be combined into a comprehensive predictive model. Once the final model is constructed, it will be used to forecast the results of games played after the cutoff date, and the model's accuracy will be evaluated against the actual outcomes recorded in the test data.

2 Exploratory Data Analysis



We used a heat map to visualize the relationships between numerical variables using color gradients, where the intensity of color indicates the strength of the correlation. This heat map shows how features like PTS, FG%, AST, and REB are interrelated. For example, PTS is strongly correlated with FG%, suggesting that better shooting accuracy contributes to higher scores. It also reveals a positive correlation between AST and PTS, reflecting how teamwork impacts scoring efficiency. The insights indicate that key performance metrics like scoring and assists significantly contribute to game outcomes, making them important predictors.



The bar graph illustrates the average values of key performance metrics for games won and lost. Winning teams generally have higher values for PTS, FG%, 3P%, and REB while maintaining a positive +/- score, indicating stronger offensive and defensive performance. Conversely, losing teams show lower averages in these metrics. This graph highlights that consistent scoring, efficient shooting, and solid rebounding are critical factors distinguishing winning from losing performances.

The histograms show the distributions of key metrics. Most features exhibit a roughly normal distribution with some skewness, indicating that outlier games with exceptionally high or low stats occur but are rare. These distributions help identify typical performance ranges and guide how to treat extreme values during modeling.

From the EDA, our team sees the importance of offensive efficiency and teamwork in determining game outcomes. Correlations and average key stats suggest that well-rounded performances across multiple metrics are associated with wins. The consistent patterns observed across distributions indicate that most games follow predictable trends, with occasional outliers that may reflect unique circumstances. These findings provide a strong foundation for building a predictive model by focusing on metrics that drive winning performances.

3. Data Processing

In the data processing stage, we checked 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.

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 **IsHomeGame**, a binary indicator specifying whether the team was playing at home or away. Streak-based features like **Consecutive_Win_General** were added to account for team momentum. Additionally, comparative metrics such as **Shooting_Differential** and **Turnover_Differential** were derived to highlight specific aspects of team strategy and execution. To further improve the model's accuracy, we incorporated advanced metrics like **eFG** (Effective Field Goal Percentage), which adjusts for the additional value of three-point shots; **FTr** (Free Throw Rate), indicating how often a team gets to the free-throw line relative to field goal attempts; and **Off_Rating** (Offensive Rating), which quantifies points scored per 100 possessions. These metrics provide deeper insights into team efficiency and offensive performance.

We will divide the dataset into training and testing sets, using February 26, 2024, as the cutoff date. The data will be split into 70% for training and 30% for testing. This approach includes data from the trading period leading up to the deadline in the training set, which enhances our model's accuracy, as significant trades are more likely to occur during this time.

4 Modeling

To increase the model's accuracy, our group attempted to mine the hidden information in the dataset. We broke the full model into six 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 match 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 created a new column for our training data (before date 2.26), which calculates each NBA team's average home win rate. To ensure completeness for testing purposes, we assigned a default win rate (e.g., 0.5) for teams with no prior home game data.
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 primary 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 successive 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 themselves. 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 Efficiency Metrics

To capture a team's overall performance trends and integrate meaningful game metrics, we engineered six key features: **Point Differential**, **Shooting Differential**, **Turnover Differential**, **Effective Field Goal Percentage (eFG%)**, **Free Throw Rate (FTr)**, and **Offensive Rating**

(Off_Rating). These metrics compare a team's performance against its opponent in terms of scoring, shooting efficiency, and turnovers.

Point Differential measures the scoring margin by subtracting the opponent's points from the team's points, reflecting dominance or struggle. Shooting Differential highlights differences in shooting accuracy, while Turnover Differential assesses ball-handling effectiveness by comparing turnover rates. Additionally, eFG% adjusts traditional field goal percentage to account for the increased value of three-point shots, providing a more accurate measure of shooting efficiency. FTr evaluates the frequency of free throw attempts relative to field goals, indicating a team's aggressiveness in drawing fouls and capitalizing on free throws. Offensive Rating standardizes points scored per 100 possessions, accounting for game pace and possession count to offer a comprehensive view of offensive efficiency. We extended these metrics by calculating their average values for each team, ensuring pre-game information availability and enhancing feature consistency. This approach allows the model to focus on broader performance patterns rather than individual game fluctuations.

4.7 Final Model

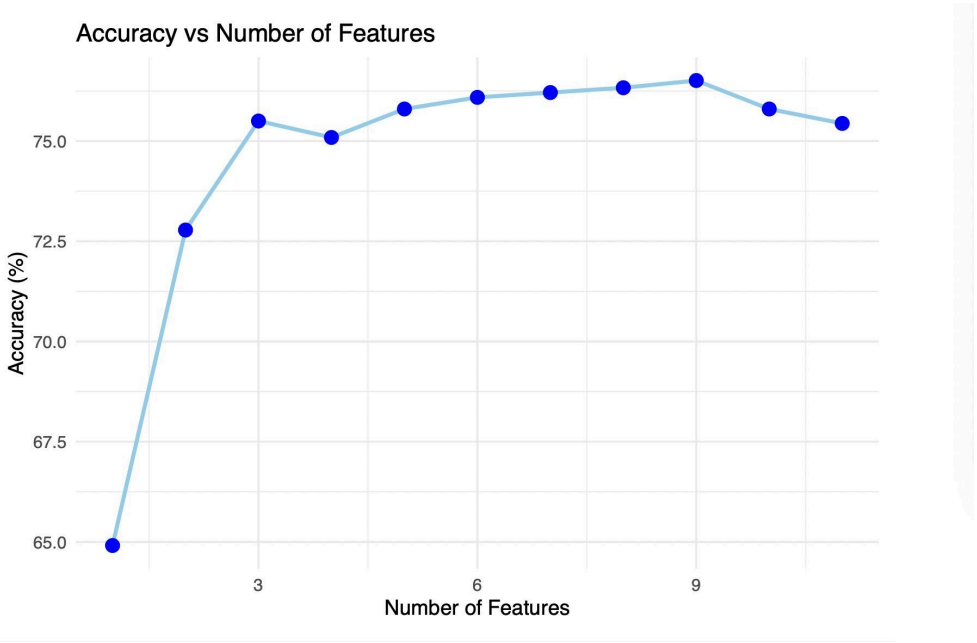
A forward selection process with k-fold cross-validation was applied to identify the best feature combinations for predicting game outcomes. The training data was divided into five folds for cross-validation. Starting with a null model, features were added iteratively based on their ability to improve average accuracy across folds. The process continued until no significant accuracy improvement was observed. This method systematically evaluated feature contributions while reducing overfitting risks through cross-validation. It identified the most effective combinations of features with a balance between model simplicity and predictive performance.

Table 1: Best Feature Combinations by Feature Count

Step	Combination	Accuracy	Num_Features
1	OR	64.91	1
2	OR, IHG	72.78	2
3	OR, IHG, HWR	75.50	3
4	OR, IHG, HWR, CWG	75.09	4
5	OR, IHG, HWR, CWG, FTr	75.80	5
6	OR, IHG, HWR, CWG, FTr, TD	76.09	6
7	OR, IHG, HWR, CWG, FTr, TD, CLG	76.21	7
8	OR, IHG, HWR, CWG, FTr, TD, CLG, eFG	76.33	8
9	OR, IHG, HWR, CWG, FTr, TD, CLG, eFG, SD	76.51	9
10	OR, IHG, HWR, CWG, FTr, TD, CLG, eFG, SD, WAP	75.80	10
11	OR, IHG, HWR, CWG, FTr, TD, CLG, eFG, SD, WAP, ASD	75.44	11

The analysis revealed that accuracy increased as features were added but plateaued and declined when excessive features were included. The highest accuracy was achieved with nine features,

emphasizing the importance of avoiding overfitting by limiting complexity. Models with fewer features performed well, demonstrating the robustness of forward selection in identifying essential predictors. (See the reference page for abbreviation)



5 Strengths 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. Moreover, our model approach has limitations. Forward selection evaluates features sequentially, which can overlook combinations where less predictive features become significant only when paired with others. Additionally, k-fold cross-validation increases computational time, particularly for datasets with many features, and may still introduce variability in accuracy estimates depending on the data split.

6 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 75.44%. 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. Overall, the model serves as a robust starting point for predictive analysis in sports.

7. Reference List

Table 2: Abbreviation Table

Abbreviation	Meaning
ASD	Avg_Score_Diff
WAP	Weighted_Avg_PTS
IHG	IsHomeGame
HWR	HomeWinRate
SD	Shooting_Differential
TD	Turnover_Differential
CWG	Consecutive_Win_General
CLG	Consecutive_Loss_General
eFG	Effective Field Goal %
FTr	Free Throw Rate