From Stats to Scores: Analyzing NBA Game Outcomes with Random Forest and QDA

1. Introduction

For this project, we conducted an in-depth analysis of the 2023-24 NBA Dataset provided to us. The NBA Dataset is an organization by a team of game statistics for each match during the 2023-24 season. Our objective was to test different models in order to predict the outcome of each game based on historical data. The dataset included features such as team name, what teams played in the match and where it occurred, game date, whether the team won or lost, points scored, how points were scored, attempted points, types of moves used, and more.

We decided to use feature engineering before testing our models with Quantitative Discriminant Analysis (QDA) and Random Forest. With our selected features, Random Forest performed the best with a testing accuracy of 78%. In this paper, we provide a detailed breakdown of our analysis, including data preprocessing, experimental setup, and results and analysis.

2. Data Preprocessing

Before we could train our models, we needed to take the necessary steps to prepare our data. We chose to make new features from the existing data, creating rolling averages for each game statistic. After finding the averages, we also constructed a separate dataset featuring the differences between the rolling averages. This was followed by selecting the significant variables that we would use to train the models and assigning weights to certain game statistics. When training our models we decided to assign higher weights to game statistics based on how recently they occurred previous to the match being considered.

2.1 Cleaning Data

Before we could do any feature engineering or analysis, we first had to clean our dataset to make sure it would work for the methods we were using. First, we checked for missing data and removed any rows that contained NA variables. Next, we converted the Win/Loss column to binary variables using 1 to represent a win and 0 to represent a loss. Similarly, we created a new column of binary variables to signify whether or not the match was a home or an away game for the corresponding team. For this column, 1 represents a home game and 0 represents an away game. Each of these steps were crucial to preventing errors when doing our analysis and ensuring our results would not be skewed in any way by the dataset.

2.2 Rolling Averages

In order to predict the outcome of the game, we thought it would be best to only look at data from before the match we are trying to predict. For this reason, we chose to use the rolling averages for each game statistic in our analysis. Rolling averages allow us to look at how a team has been performing in more recent games rather than relying on statistics from the entire season. This prevents our assessment of a team's current capabilities from being influenced by the outcomes of previous games that may no longer accurately reflect the team's performance.

2.3 Differences

After creating the rolling averages, we created a separate dataset of the differences in the averaged game statistics. Taking the difference in the rolling averages between the team and their opponent would allow us to compare the performances of both teams leading up to the game. Rather than analyzing each team's averages separately, this information highlights the strengths and weaknesses of each team relative to their opponent. Using the differences in our analysis allows our model to take into account the interplay between the recent trends of both team's improving its ability to recognize key factors that affect the outcome of a match.

3. Experimental Setup

We evaluated the models through two main approaches: Quadratic Discriminant Analysis (QDA) and Random Forest.

- **QDA**: We decided to use QDA to leverage its ability to create flexible decision boundaries by assuming a joint normal probability distribution for each class. We trained the model on the training dataset and evaluated its performance on the test set using accuracy, a confusion matrix, and a classification report.
- **Random Forest**: Random forest is a method that constructs estimates from random subsets of the data based upon the majority vote of the trees for classification. We split the dataset into 70% training and 30% testing subsets and evaluated the model using accuracy and classification metrics.

4. Results and Analysis

To evaluate the effectiveness of our predictive models, we have implemented and compared two approaches: QDA and Random Forest. We assessed these models through four different datasets that are derived from our feature engineering process: rolling averages, weighted rolling averages, and difference-based rolling averages.

4.1 Rolling Averages

The first dataset we tested focused on rolling averages of the game statistics from the game. This approach has aimed to capture trends in team performances by averaging all of the metrics from the previous games up to the games that are ready to be predicted.

- Random Forest has achieved an accuracy of approximately 59%. This is only slightly outperforming the guessing and has suggested that simple rolling averages from the dataset itself are not sufficiently enough to account for the complexity of the game
- QDA has resulted in a similar accuracy of around 57% in which it slightly underperformed compared to the random forest. This might be cause that QDA is more effective for datasets that contain Gaussian distribution. The inherent non-linearity and interactions in basketball would perhaps make it less suitable.

4.2 Weighted Rolling Averages

To improve the prediction from the random forests, we have developed weighted rolling averages. This average will prioritize more recent games by assigning a higher weight to their statistics. This adjustment has quickly demonstrated the fact that recent performance is more indicative of the team's winning probability compared to the older games which are less applicable.

- Random forests have performed significantly better with this dataset with an accuracy of around 78%. This improvement from the original 59% has demonstrated its prediction power.
- QDA on the other hand has also shown improvement from 57% to 75%. However, though, it did not surpass the prediction accuracy for Random Forest once again.

4.3 Difference-Based Rolling Averages

We continue to explore the dataset using differences in rolling averages. This approach will account for the relative performances of a team compared to its opponent by taking the differences between their respective rolling averages. By focusing more on the interplay between teams, we hope to find out whether or not the performance of the team compared to each other will matter in their success.

- With this dataset, the Random Forest model has achieved an accuracy of 62%, a drop compared to weighted rolling but a rise compared to the rolling average.
- QDA achieved an accuracy of 59%, once again underperforming Random Forest. However, compared to the rolling averages, it has slightly improved from 57% but it is still hard to say whether or not 2% is a huge increase.

4.4 Comparison of Models

From our analysis, we can see the Random Forest model has been consistently outperforming QDA through all three datasets which makes sense because it has more versatility and the ability to model complex relationships in data. It is also important to note that by weighing the data, we can see the change of accuracy significantly. This could be because of the constant change in rosters in the NBA team.

4.5 Feature Importance

Lastly, to conclude our research, we have performed feature importance through forward selection in which W/L - Weighted Rolling Avg, DREB - Weighted Rolling Avg, and 3PA - Weighted Rolling Avg were the most significant variables that influenced the dataset.