Decoding NBA Playoff Qualification: A Statistical Analysis

PSTAT 232 Final Project

Elijah Castro, Sarkis Der Wartanian, Cyrus Tabatabai-Yazdi

March 20, 2024

Abstract

Our study delves into the intricate relationship between team performance metrics and playoff success in the National Basketball Association (NBA) using statistical modeling and machine learning techniques. We first identify the key determinants driving a team's qualification for the postseason by leveraging the Random Forest algorithm to select the most influential features from a pool of team statistics. Subsequently, we construct a Neural Network model to further refine our understanding, capturing complex patterns in the data. The models are evaluated through cross-validation to ensure their generalizability to unseen data. Our findings suggest that factors such as opponent field goal percentage, points per game (both team and opponent), and team field goal percentage play crucial roles in playoff qualification. We also explore the performance of different optimization methods in the Neural Network model, highlighting the effectiveness of Adamax and Adam optimizers.

Contents

Introduction

In the world of professional basketball, the National Basketball Association (NBA) stands as a beacon of excellence, showcasing the talents of the finest players and teams. With its thrilling games and passionate fan base, the NBA has captured the hearts of millions worldwide, becoming a symbol of athleticism and sportsmanship.

Behind the glitz and glamour of the NBA lies a fiercely competitive landscape, where teams battle relentlessly throughout the season for a chance to compete in the ultimate stage of basketball excellence – the NBA playoffs. The playoffs represent the culmination of months of dedication, perseverance, and strategic prowess, where only the best of the best earn the right to contend for the coveted NBA championship.

The journey to playoff qualification is a complex and intriguing one, influenced by a myriad of factors that extend beyond mere win-loss records. Team statistics, such as points per game, rebounds, assists, and defensive efficiency, play a crucial role in determining a team's success throughout the season and, ultimately, their playoff fate.

The NBA data analyzed spans the seasons from 1980 to 2023, encompassing a broad array of team and opponent statistics. The following provides a succinct description of each unique statistic considered in our analysis:

Table 1: Variable Notation and Description

Variable	**Description**
`year`	Year of NBA Season
`Team`	Team Name
`playoffs`	Playoffs Indicator (1 if team made playoffs, 0 otherwise)
Team Performance	
`Rk_Team`	Ranking of Offensive Rating - Points Scored per Game
`FG_Team`	Field Goals Made per Game, Attempted per Game, and Percentage
`3P_Team`	3-Point Field Goals Made per Game, Attempted per Game, and Percentage
`2P_Team`	2-Point Field Goals Made per Game, Attempted per Game, and Percentage
`FT_Team`	Free Throws Made per Game, Attempted per Game, and Percentage
`ORB_Team`	Offensive Rebounds per Game
`DRB_Team`	Defensive Rebounds per Game
`TRB_Team`	Total Rebounds per Game
`AST_Team`	Assists per Game
`STL_Team`	Steals per Game
`BLK_Team`	Blocks per Game
`TOV_Team`	Turnovers per Game
`PF_Team`	Personal Fouls per Game
`PTS_Team`	Points per Game
Opponent Performance	
`Rk_Opp`	Ranking of Defensive Rating - Points Allowed per Game
`FG_Opp`	Field Goals Made by Opponents per Game, Attempted per Game, and Percentage
`3P_Opp`	3-Point Field Goals Made by Opponents per Game, Attempted per Game, and Percentage
`2P_Opp`	2-Point Field Goals Made by Opponents per Game, Attempted per Game, and Percentage
`FT_Opp`	Free Throws Made by Opponents per Game, Attempted per Game, and Percentage
`ORB_Opp`	Offensive Rebounds by Opponents per Game
`DRB_Opp`	Defensive Rebounds by Opponents per Game
`TRB_Opp`	Total Rebounds by Opponents per Game
`AST_Opp`	Assists by Opponents per Game
`STL_Opp`	Steals by Opponents per Game
`BLK_Opp`	Blocks by Opponents per Game
`TOV_Opp`	Turnovers by Opponents per Game
`PF_Opp`	Personal Fouls by Opponents per Game
`PTS_Opp`	Points by Opponents per Game

In this project, we delve deep into the realm of NBA playoff qualification, aiming to uncover the intricate relationship between team per game statistics and postseason success. By analyzing the standardized values of these statistics and employing advanced computational statistical and machine learning techniques, we seek to identify the key factors that significantly impact a team's chances of making it to the playoffs.

Exploratory Data Analysis (EDA)

In this section, we will focus on visualizing key team performance metrics over the years to understand trends and patterns related to playoff qualification. The analysis will include average ratings for playoff teams, average team and opponent points per game, average 3-point attempts per game, and average blocks per game.

Average Offensive/Defensive Rating Rankings (Playoff Teams)

To start our analysis, we examined the average offensive and defensive rating rankings for teams that make the playoffs from 1980 to 2023. These rankings indicate where each team stands in terms of offensive and defensive performance compared to the rest of the NBA for each season.

- **Average Offensive Rating Rank**: This statistic represents the average points scored per game by a team, indicating their offensive efficiency and scoring prowess. A lower rank in offensive rating suggests that a team is more effective at scoring points against their opponents.
- **Average Defensive Rating Rank**: This statistic represents the average points allowed per game by a team, showcasing their defensive efficiency and ability to limit their opponents' scoring. A lower rank in defensive rating indicates that a team is better at preventing their opponents from scoring.

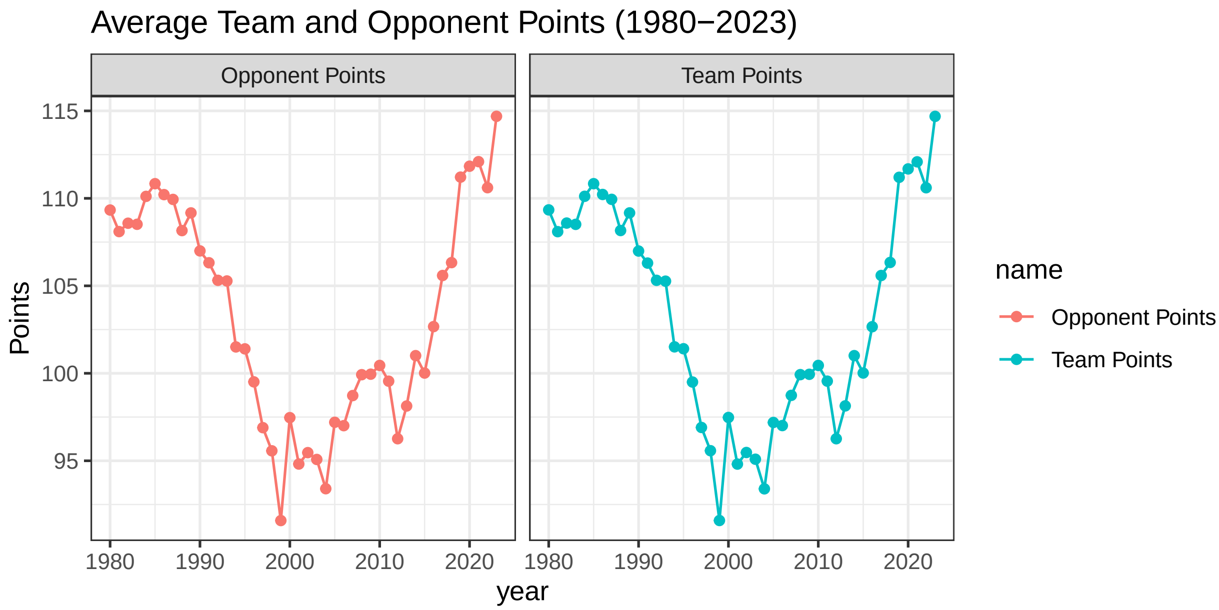
Table 2: Average Offensive and Defensive Rating Ranks for Playoff Teams (1980-2023)

Statistic	**Average Rank**
Average Offensive Rating Rank	11.857
Average Defensive Rating Rank	10.983

The table reveals that teams aiming for the postseason must achieve offensive and defensive rankings superior to the average NBA team, which holds a ranking of 15 (due to the league's 30 teams). Both offensive and defensive rankings typically fall around 11-12 for playoff-bound teams. This aligns with expectations, as playoff teams are expected to outperform the league average, ensuring their rankings are higher for both offensive and defensive metrics.

Average Team and Opponent Points Per Game

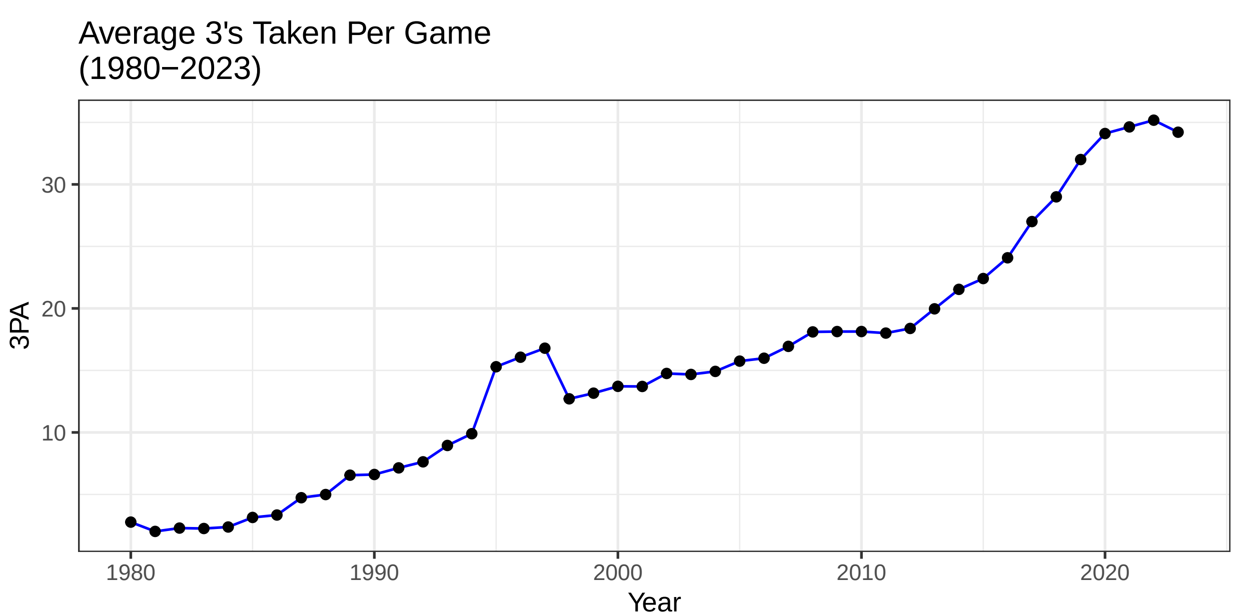
Next, we examined the average points scored by teams and points scored by opponents per game over the seasons from 1980 to 2023. The graph below illustrates how these values have evolved over the years:



Initially, from 1980 to 1999, there was a noticeable decrease in both team and opponent points scored per game. However, since 2020 to the present, there has been a general increase in both team and opponent points per game over the years, indicating a trend towards higher-scoring games. This trend may imply changes in offensive strategies, player performance, or rule adjustments that have influenced the scoring dynamics of NBA games.

Evolution of 3-Point Attempts Per Game

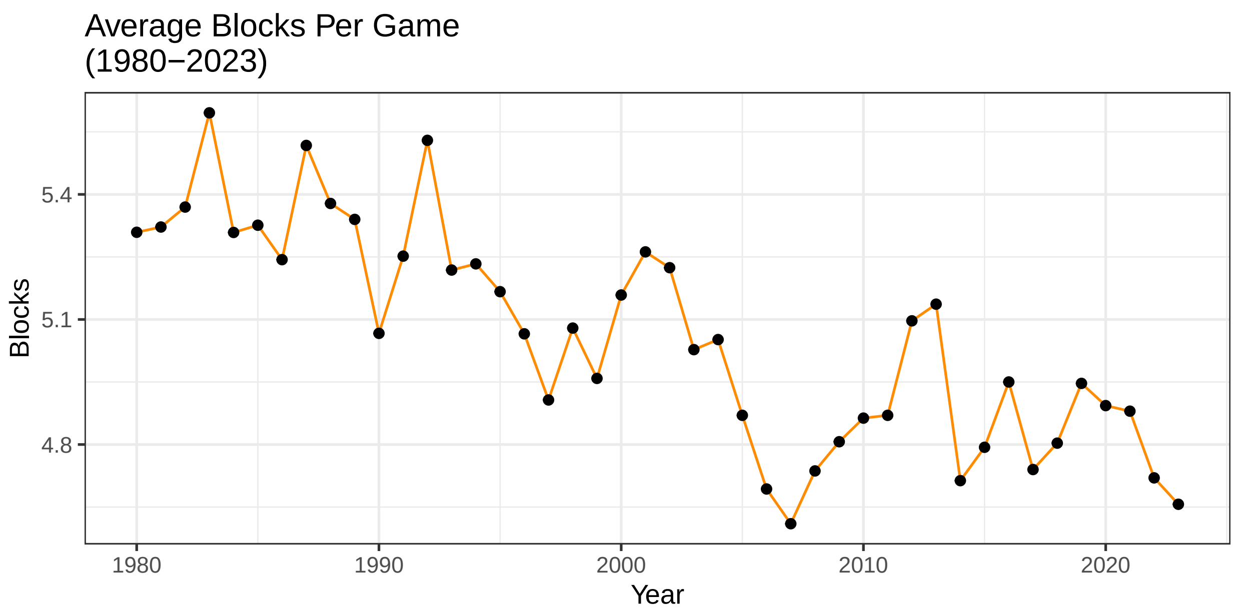
The evolution of the game of basketball is also reflected in the increasing emphasis on 3-point shooting. The graph below shows the average 3-point attempts per game by teams from 1980 to 2023:



The graph illustrates a significant rise in the average number of 3-point attempts per game over the years, highlighting the growing importance of the 3-point shot in modern basketball strategies. This trend underscores the shift towards a more perimeter-oriented style of play in the NBA.

Average Blocks Per Game

Finally, we examined the average blocks per game by teams over the seasons from 1980 to 2023. The graph below illustrates the trend in average blocks per game:



The graph illustrates minor fluctuations in average blocks per game over the years, with some seasons exhibiting higher averages than others. Nonetheless, these fluctuations have generally shown a slight decrease over time. This trend might be attributed to changes in defensive strategies, player skill levels, or rule modifications that influence shot-blocking opportunities.

Overall, these visualizations provide valuable insights into the evolving nature of the NBA game, highlighting key trends and patterns related to team performance metrics.

Modeling and Evaluation

With a more comprehensive understanding of the historical trends and dynamics within the NBA, we transition to the core of our analysis: statistical modeling. Our aim is to pinpoint the key determinants that drive a team's qualification for the postseason. By leveraging advanced techniques in feature selection and machine learning, we seek to unravel the intricate relationship between team performance metrics and playoff success.

Our modeling approach involves two crucial steps. First, we employ the **Random Forest** algorithm to identify the most influential features from a pool of team statistics. This step helps us narrow down our focus to the variables that have the greatest impact on playoff qualification. Subsequently, we construct a

Neural Network model, a powerful tool capable of capturing complex patterns in the data. Within this step, we integrate cross-validation to ensure our model generalizes well to unseen data and provides reliable predictions, further refining our understanding of the relationship between team performance metrics and playoff success.

Feature Important Analysis with Random Forest

Random Forest is a versatile and widely used ensemble learning technique for classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the average prediction (regression) of the individual trees. In our context, where we aim to identify the most influential features for playoff qualification (a classification problem) from a pool of team statistics, we implement the Random Forest algorithm as follows:

Table 3: Algorithm Steps

Algorithm 1: Random Forest
Require:
Training dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$,
number of trees T, number of features to consider at each split m
Do:
1. Initialize an empty forest F
2. For $t = 1$ to T:
a. Sample a bootstrap sample D_t from D (size n with replacement)
b. Select a random subset of features of size m
c. Grow a decision tree T_t from D_t using the selected features:
i. For each node of T_t:
- Randomly select m features without replacement
- Calculate the best split using one of the impurity measures (e.g., Gini impurity, entropy)
- Split the node based on the best split
ii. Continue until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf)
d. Add T_t to the forest F
3. Return the trained Random Forest model F

Model Building

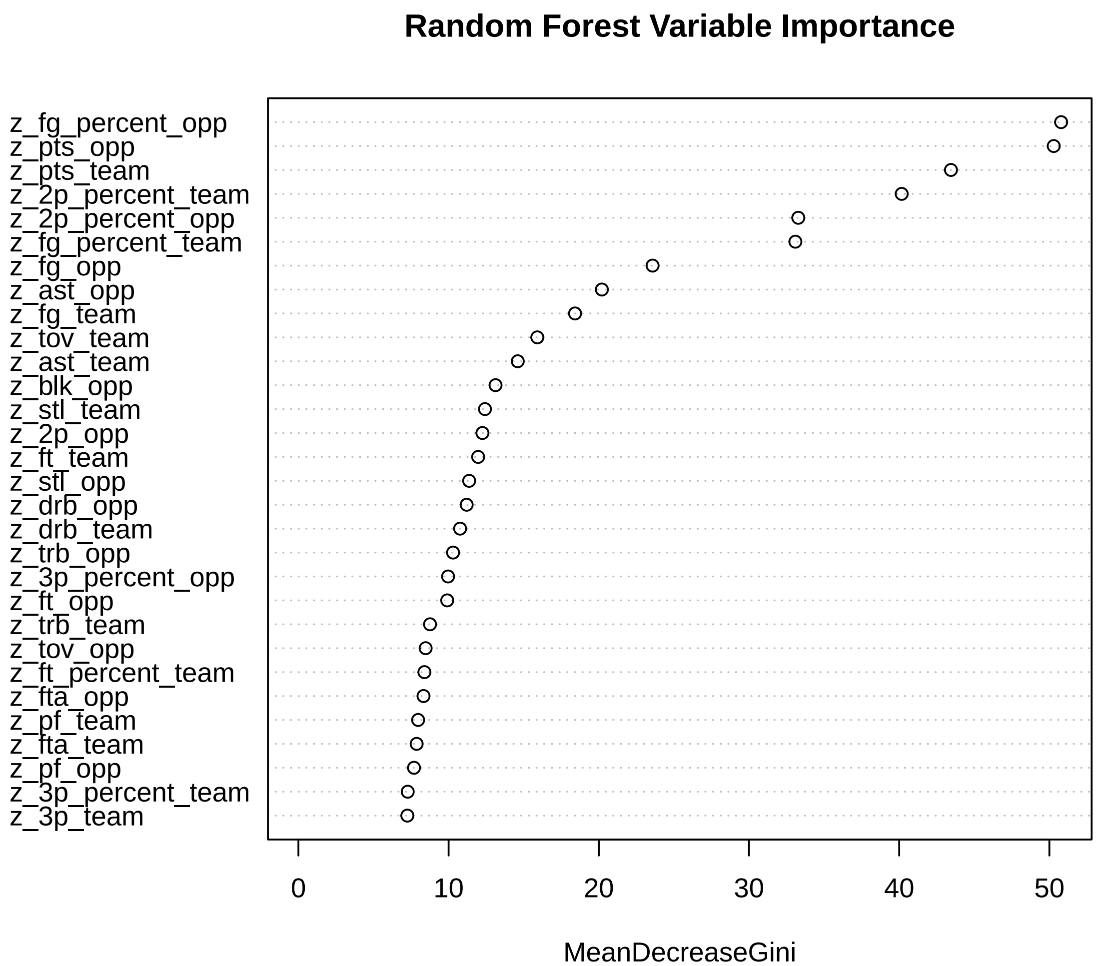
The process begins by standardizing the data using z-scores, which transforms the variables to have a mean of 0 and a standard deviation of 1, ensuring that all variables are on the same scale. This step is crucial for models like Random Forest, which are sensitive to the scale of the input features.

Next, the features needed for the analysis are selected and cleaned. This includes both offensive and defensive team statistics, such as points, field goal percentage, and rebounds, which are considered in both team and opponent contexts to capture the full spectrum of performance metrics.

The Random Forest algorithm is then employed on the preprocessed data. This algorithm builds numerous decision trees using various subsets of the data and features. Each decision tree is trained to predict a team's playoff qualification based on the selected features. The final prediction is made by combining the predictions of all trees, using the mode since this is a classification problem. (Liaw and Wiener 2002)

Results and Interpretations

The Random Forest model identified several key features that significantly influence a team's qualification for the NBA playoffs. These insights are visually confirmed in the feature importance plot below, which provides a clear overview of the relative importance of each feature in the model's decision-making process. The plot highlights the impact of the following features, ranked by their decrease in Gini impurity (`MeanDecreaseGini`): (Kuhn 2022)

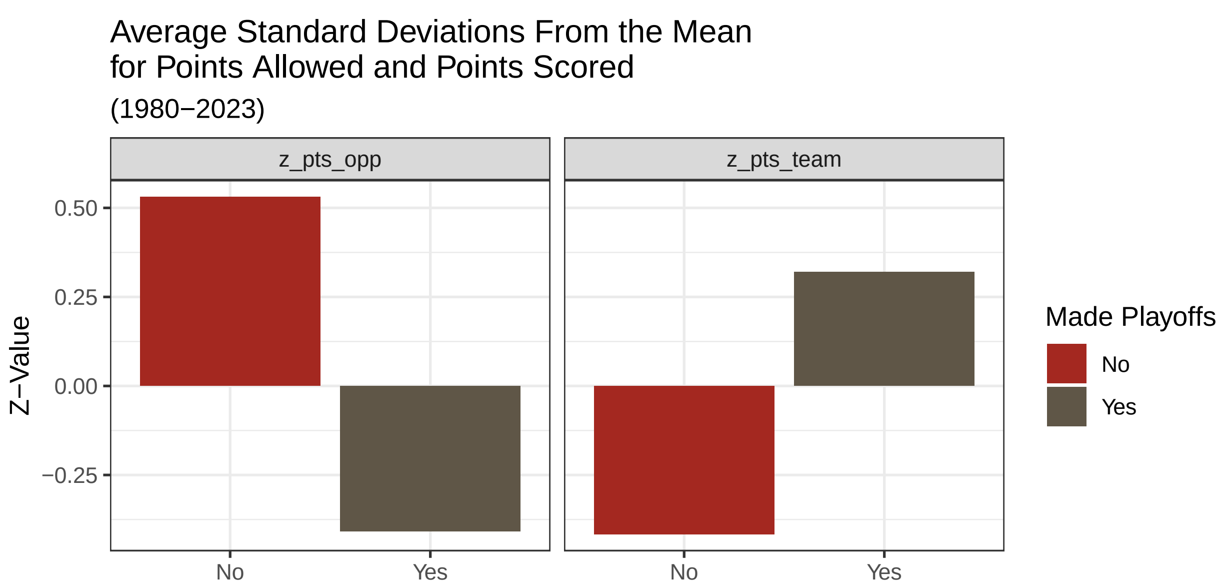


- **Opponent Field Goal Percentage** (`z_fg_percent_opp`): This feature has the highest importance, suggesting that a team's defensive performance, particularly in limiting their opponent's field goal percentage, plays a crucial role in playoff qualification.
- **Opponent Points Per Game** (`z_pts_opp`): Similar to team points per game, the number of points allowed to opponents is an important indicator of a team's defensive capabilities and overall performance.
- **Team Points Per Game** (`z_pts_team`): The average number of points scored by a team per game is another critical factor. Teams that score more points tend to have a higher likelihood of making it to the playoffs.

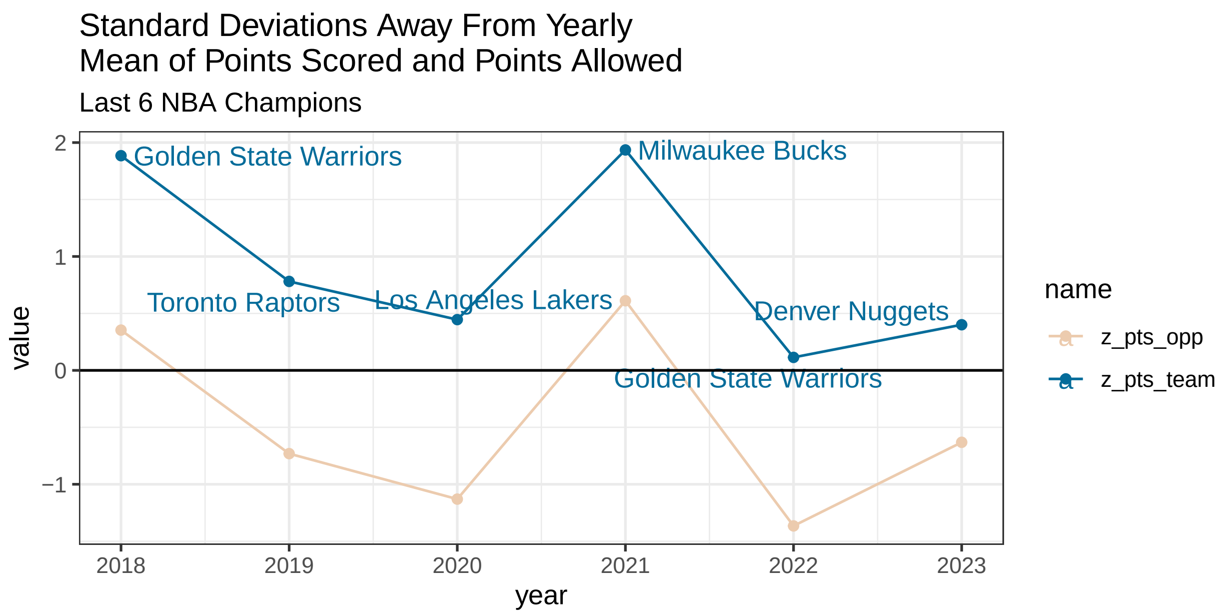
- **Team 2-Point Field Goal Percentage** (`z_2p_percent_team`): The efficiency of a team's 2-point field goal attempts is also a significant factor. Teams with a higher 2-point field goal percentage tend to have better offensive performance, contributing to their playoff success.
- **Opponent 2-Point Field Goal Percentage** (`z_2p_percent_opp`): Similar to the team's 2-point field goal percentage, limiting the opponent's efficiency in 2-point field goals is crucial for defensive success and playoff qualification.
- **Team Field Goal Percentage** (`z_fg_percent_team`): The overall field goal percentage of a team is another important factor. A higher field goal percentage indicates better shooting efficiency, which is a key aspect of offensive performance.

Having identified the most influential features through the Random Forest model, we now delve into visual representations of these key metrics. (Ram and Wickham 2023) These visualizations provide a deeper understanding of how certain performance indicators impact a team's success in making the playoffs.

The first visualization focuses on the average standard deviations from the mean for points allowed and points made. It reveals a stark difference between teams that made the playoffs and those that did not. For points allowed, playoff-bound teams were consistently below the mean, indicating stronger defensive capabilities. Conversely, non-playoff teams tended to allow more points, often above the mean. This suggests that a strong defense is crucial for playoff qualification. For points made, playoff teams tended to score more points than non-playoff teams, highlighting the importance of offensive prowess in securing a playoff spot.



The second visualization examines the standard deviations of points scored and allowed for the last six NBA champions. It showcases the performance dynamics of these championship teams, highlighting their offensive and defensive strategies. Notably, teams like the Warriors in 2018 and Bucks in 2021 had the highest standard deviations above the mean for points scored but also, the only two teams with standard deviations *above* the mean for points allowed, indicating a more offensive focused approach. On the other hand, teams like the Lakers in 2020 and Warriors in 2022 had smaller standard deviations above the yearly mean for points scored but significantly larger standard deviations *below* the mean for points allowed, showcasing their strong defensive strategies. (Firke 2023)



These visualizations underscore the importance of both offensive and defensive performance in NBA teams' quest for playoff qualification and championship success. By understanding and leveraging these key features, teams can potentially enhance their performance and increase their chances of qualifying for the NBA playoffs.

Predictive Model with Neural Networks

Neural networks are a class of machine learning algorithms inspired by the structure and function of the human brain. They consist of interconnected nodes, or "neurons," organized in layers. Each neuron receives input, processes it through an activation function, and passes the output to the next layer. Neural networks are capable of learning complex patterns in data and are widely used for tasks such as classification and regression.

Keras is a high-level neural networks API written in Python that runs on top of TensorFlow. It provides a user-friendly interface for building and training neural networks, making it easier to experiment with different architectures and configurations.

Model Architecture

For our predictive model, we construct a neural network with the following structure:

- **Input Layer**: 6 units, corresponding to the 6 selected features (e.g., team/opponent points, field goal percentage, 2-point field goal percentage).

- **Hidden Layers**:
  - First hidden layer with 128 units and ReLU activation function.
  - Second hidden layer with 64 units and ReLU activation function.
- **Output Layer**: 2 units (binary classification), with a sigmoid activation function. The model is compiled using the binary cross-entropy loss function and the accuracy metric, and the optimizer is specified based on the specific experiment (e.g., SGD, RMSprop, Adam, etc…).

Cross-Validation

To ensure the robustness of our model, we perform cross-validation. This technique involves splitting the data into k subsets (folds), training the model on k-1 folds, and evaluating it on the remaining fold. This process is repeated k times, with each fold serving as the validation set exactly once.

We implement cross-validation within the context of our neural network modeling. We split the data into 5 folds and train the model on 4 folds while using the remaining fold for validation. This process is repeated 5 times, each time using a different fold for validation. This approach allows us to assess the model's performance across different subsets of the data and helps prevent overfitting.

Optimizers

We compare the performance of the model using different optimizers, including: (Allaire and Chollet 2022)

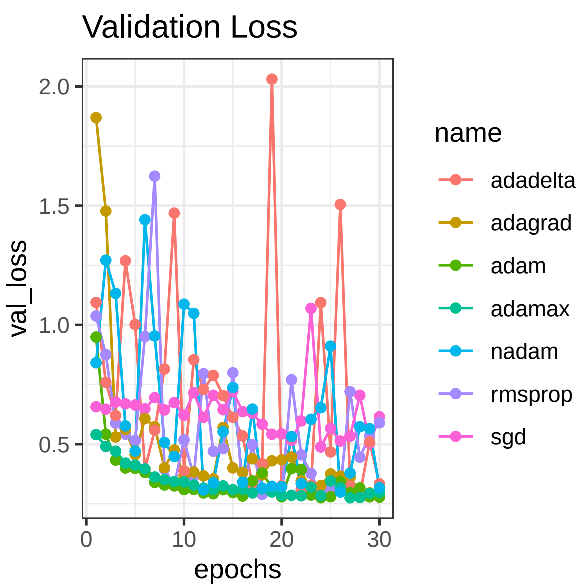
- **SGD (Stochastic Gradient Descent)**: A classic optimizer that updates the weights of the network based on the gradient of the loss function with respect to the weights. It is called "stochastic" because it randomly selects a subset of training samples (mini-batch) to compute the gradient, making it computationally efficient.
- **RMSprop (Root Mean Square Propagation)**: An adaptive learning rate method that adjusts the learning rate for each parameter based on the average of recent gradients. It divides the learning rate by the square root of the exponentially weighted average of squared gradients.
- **Adagrad (Adaptive Gradient Algorithm)**: An adaptive learning rate method that adapts the learning rate for each parameter based on the history of gradients for that parameter. It scales the learning rate inversely proportional to the square root of the sum of the squared gradients.
- **Adadelta**: An extension of Adagrad that addresses its diminishing learning rates by restricting the accumulation of past gradients to a fixed-size window of the most recent gradients.
- **Adam (Adaptive Moment Estimation)**: Another adaptive learning rate method that combines the advantages of both RMSprop and momentum optimization. It computes adaptive learning rates for each parameter based on estimates of first and second moments of the gradients.
- **Adamax**: Avariant of Adam that uses the infinity norm (maximum absolute value) of the gradients in place of the second moment. This simplifies the algorithm and can be more efficient for some problems.
- **Nadam**: Another variant of Adam that incorporates the Nesterov accelerated gradient (NAG) method. It combines the advantages of Adam with NAG, resulting in faster convergence and better performance on some problems.

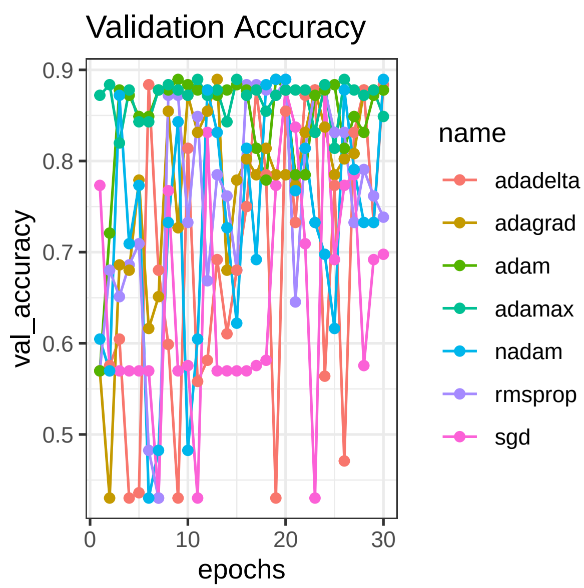
Each optimizer has its strengths and weaknesses, and we will experiment with the different optimizers to identify which one that works best for our particular neural network model.

Results and Interpretations

We visualize the validation loss and accuracy over the training epochs for each optimizer. This allows us to compare how quickly each optimizer converges and how well it generalizes to unseen data. (Pedersen 2024)

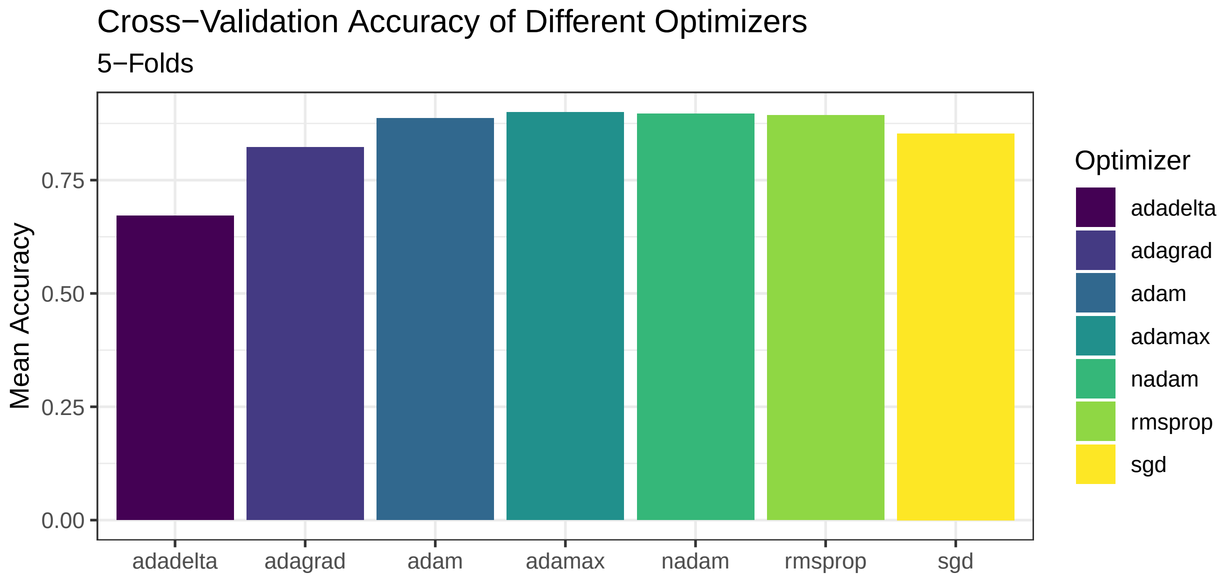
We find that the Adam-related optimizers, including Adam and Adamax, as well as Adagrad, outperformed the other optimizers in terms of validation loss and accuracy. These optimizers consistently achieved lower validation loss, around 0.5, compared to the other optimizers, which had validation loss in the 1's and 2's. Additionally, the Adam-related optimizers and Adagrad demonstrated higher validation accuracy, ranging from 0.8 to 0.9, while the other optimizers had validation accuracy in the 0.4 to 0.7 range. This indicates that Adam, Adamax, and Adagrad were more effective in optimizing the neural network model for playoff qualification prediction, converging more quickly and generalizing better to unseen data.





Additionally, we plot the mean accuracy over 5 folds from cross-validation for each optimizer to provide a summary of their performance. This metric indicates the average accuracy of the model across different splits of the data, providing a measure of how well the model generalizes to unseen data. Below, we find that Adamax performed the best, with the highest mean accuracy, followed closely by Adam. Adadelta had by far, the lowest mean accuracy.

Additionally, we plot the mean accuracy over 5 folds from cross-validation for each optimizer to summarize their performance. This metric indicates the average accuracy of the model across different splits of the data, providing insight into how well the model generalizes to unseen data. Our analysis revealed that Adamax achieved the highest mean accuracy, closely followed by Adam, while Adadelta exhibited the lowest mean accuracy by a significant margin.



Overall, this section has provided a comprehensive analysis of the key determinants of playoff qualification in the NBA. Through the use of advanced statistical techniques, including Random Forest for feature selection and Neural Networks for predictive modeling, we have identified crucial team performance metrics that significantly impact playoff success. The rigorous evaluation of our models through cross-validation has ensured their reliability and generalizability to unseen data.

Our results have revealed the importance of defensive performance, as indicated by metrics such as opponent field goal percentage and points allowed per game, in determining playoff qualification. Additionally, offensive efficiency, as reflected in metrics like team points per game and field goal percentage, has emerged as a critical factor for postseason success.

The comparison of different optimizers in our Neural Network model has further enhanced our understanding, with Adam and Adamax demonstrating superior performance in predicting playoff qualification based on the selected features.

Conclusion

In conclusion, our study sheds light on the factors influencing NBA playoff qualification, though with certain limitations. The data used for our research problem included only simple per-game statistics, limiting our ability to perform further analysis on more advanced analytics. Future work could include incorporating additional features such as the average age of players, number of years of playoff experience, coach playoff appearances, experience of the coach, average salary, and maximum salary of the roster, among others, to gain a more comprehensive understanding of playoff qualification dynamics.

Our analysis also highlights the ongoing debate in the NBA community regarding the importance of defense versus offense. While historically, defense has been considered crucial for playoff success, recent trends suggest a shift towards a more offense-oriented approach. The game has evolved significantly, with teams placing greater emphasis on three-point attempts and shooting efficiency. Despite these trends, anomalies like the Miami Heat's success in the 2023 season, despite being the worst offensive team in terms of points per game, demonstrate the multifaceted nature of playoff qualification.

The primary focus of our project was to explore the application of deep learning and different optimization methods in the context of NBA playoff qualification. By employing techniques such as Random Forests and Neural Networks, along with cross-validation, we aimed to uncover insights that could inform decision-making in professional basketball. While our study provides valuable insights, further research using more comprehensive datasets and advanced analytics could enhance our understanding of the complex dynamics underlying playoff qualification in the NBA.

References

Allaire, JJ, and Francois Chollet. 2022. *Keras: R Interface to 'Keras'*. https://CRAN.R-project.org/package=keras.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Kuhn, Max. 2022. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Ram, Karthik, and Hadley Wickham. 2023. *Wesanderson: A Wes Anderson Palette Generator*. https://CRAN.R-project.org/package=wesanderson.