**NBA Playoff Predictive Analysis**

Preston Weaver  
Professor Joe Faith, advisor  
MATH 440 Senior Project

abstract

This paper analyzes the NBA playoffs to identify key trends among past champions, aiming to improve predictions of future winners. Specifically, it examines how regular-season statistics can be used to forecast a team's postseason success.

1. Introduction

The NBA playoffs feature sixteen teams, the top eight teams from the Eastern Conference and the top eight teams from the Western Conference. Throughout an NBA season, all thirty teams play eighty-two regular season games. The winning percentage of these regular season games is used as the criteria for playoff qualification. In each conference, the eight teams with the highest winning percentages qualify for the playoffs [4].

The teams are then placed in a sixteen-team bracket, in which the one seed plays the eight seed, the two seed plays the seven seed, the three seed plays the six seed, and the four seed plays the five seed. This occurs in both conferences. Each matchup plays a best-of-seven series, the winner of which moves on to the next round. After 4 rounds, one team is left standing and crowned the NBA Champion.

In 2021, the three-seeded Milwaukee Bucks were crowned the Champs after defeating the Phoenix Suns. This surprised many, as some sports betting organizations gave the Bucks the sixth best odds of winning the Championship [3]. A three seed ultimately winning the championship demonstrates that regular season win percentage alone does not fully predict post season success. This raises key questions: Can we predict which of the sixteen playoff teams is most likely to win the championship? And which regular-season metrics are the best indicators of postseason success? This paper uses machine learning to explore those answers.

**2. DATA PREPARATION**

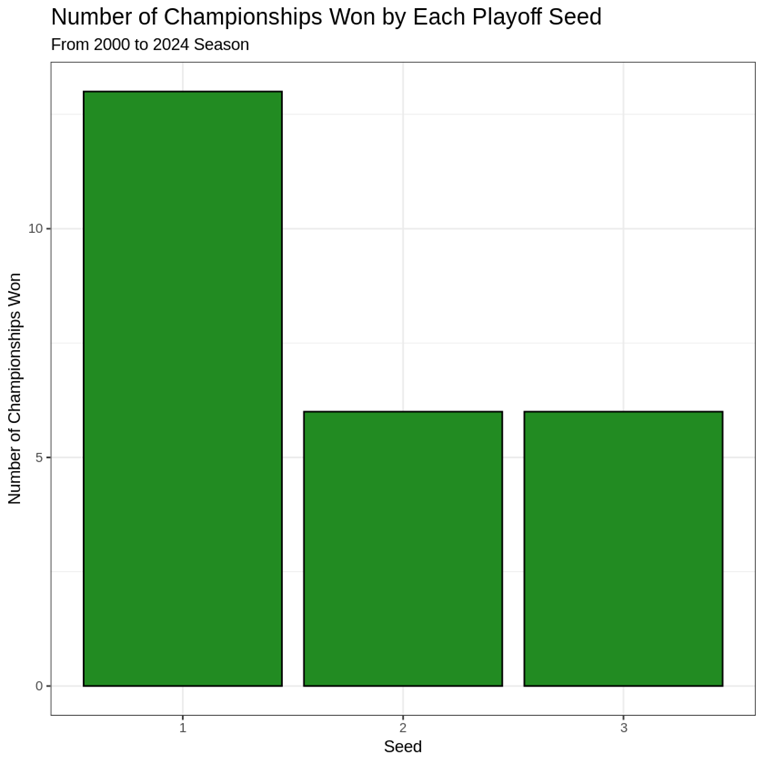
This analysis aims to predict postseason wins using regular-season statistics. To achieve this, regular-season data was collected from Basketball Reference [1], including offensive, defensive, and advanced metrics for all teams from the 2000 to 2024 seasons. Postseason data—comprising the season, team, and playoff wins—was compiled into a separate dataset. These two datasets were then merged into a single data frame, forming the basis for the analysis.

An additional step needed to be taken, however. As many avid NBA fans know, the game of basketball has changed quite a bit in the past twenty-five years. This is evident in various statistics, but for illustration, consider points per game (PPG). In the 2000 season, the league average for PPG was 97.5, with the highest-scoring team averaging 105 PPG. By the 2024 season, the league average had risen to 114 PPG, and even the lowest-scoring team averaged 105.8 PPG. In other words, the worst-scoring team in 2024 would have been the best-scoring team in 2000. Given this shift, an adjustment was needed to account for the changes in the statistical landscape of the NBA.

To account for these changes, a Z-score transformation was applied to the data. Each statistic was standardized within its season by subtracting the season’s mean and dividing by the season’s standard deviation. This process ensured that all statistics are contextualized relative to their respective season, allowing for more meaningful comparisons across years.

3. Exploratory Data Analysis

Before modeling began, exploratory data analysis (EDA) was conducted to identify key predictor variables. While many potential predictors were considered, this section highlights a few of the most impactful. One of the strongest predictors identified was playoff seed. Over the past twenty-five seasons, the No. 1 seed has been the most common champion. Notably, no team seeded lower than No. 3 has won the championship in this span. Of the last 25 champions, 13 were No. 1 seeds, six were No. 2 seeds, and six were No. 3 seeds.



**Figure 1.** Bar chart of number of championships won by each seed

Another significant predictor identified was the presence of a top-ten scorer which was defined as a player ranking among the top ten in points per game (PPG) for that season. Of the last twenty-five champions, seventeen had at least one top-ten scorer on their roster, while only eight did not. This suggests that having a dominant scorer is a strong indicator of playoff success.

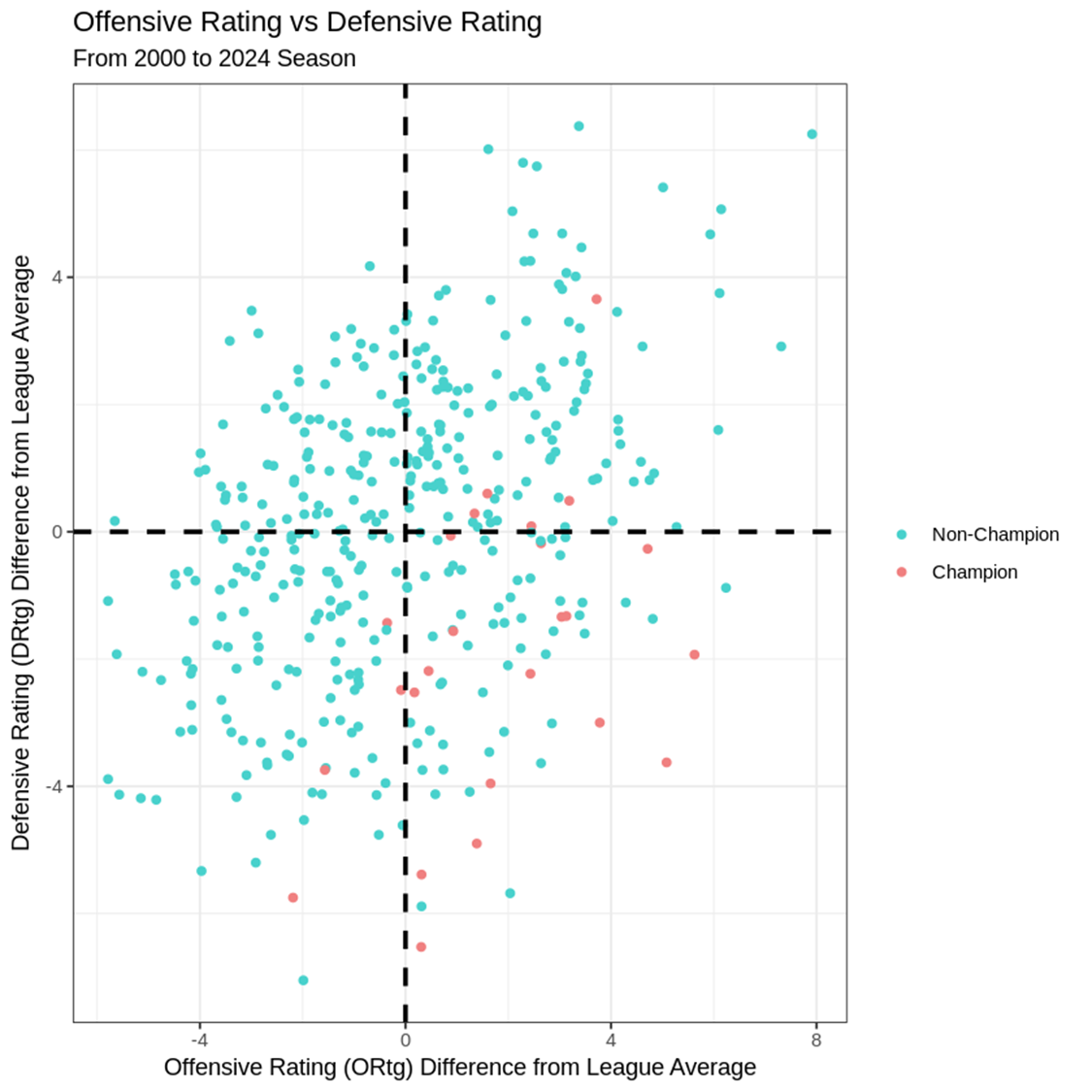
**A graph of a number of scores

AI-generated content may be incorrect.**

**Figure 2.** Bar chart of championships won by teams with top 10 scorers

Additional key predictors identified were offensive rating (ORtg) and defensive rating (DRtg). ORtg is an advanced statistic that measures a team's offensive efficiency, representing the number of points a team is expected to score per 100 possessions. Conversely, DRtg evaluates a team's defensive efficiency, representing the number of points a team is expected to allow per 100 possessions.

To examine their combined impact, a scatterplot was created. The x-axis represents the team's ORtg relative to the league average for that season, while the y-axis represents the team's DRtg relative to the league average. Positive ORtg values indicate above-average offenses, while negative DRtg values indicate above-average defenses. As shown in Figure 3, the majority of champions (64%) excelled on both ends of the court, possessing both an above-average offense and defense.



**Figure 3.** Scatterplot of Offensive Rating vs Defensive Rating

4. model Creation

After the data had been prepared and explored, the project moved to the modeling phase. The model developed was designed to take regular season statistics as predictors and estimate the number of playoff games won. The goal was to achieve a root mean squared error (RMSE) below four when evaluated on the test set. This goal was chosen because four wins are required to win a series, so an average error of less than one series win was used as a benchmark. Several models were tested, and key results are highlighted below.

**4.1. Multiple Linear Regression**

The first model tested for this project was a linear regression model. To generate a final model, backwards elimination was used to balance model performance while keeping only statistically significant predictors. The final model produced an R2 value of 0.517 meaning the model could explain a little over 50% of the variance in the response variable. When this model was tested on an independent test set, the linear model produced a root mean squared error (RMSE) of 4.049. Several predictors were deemed significant, three of the most significant predictors were win percentage, net rating, and SRS. Below is a table of how the model performed on the 2024 post season.

**Table 1.** Comparison of Linear Model Predictions to 2024 Post Season

|  |  |  |  |
| --- | --- | --- | --- |
| ***Seed*** | ***Team*** | ***Wins*** | ***Win Predictions*** |
| 1 | BOS | 16 | 14.2 |
| 1 | OKC | 6 | 9.7 |
| 2 | DEN | 7 | 8.8 |
| 3 | MIN | 9 | 7.2 |
| 3 | MIL | 2 | 6.8 |
| 6 | PHX | 0 | 6.6 |
| 8 | NOLA | 0 | 6.6 |

The model correctly predicted the Boston Celtics to win the championship, but the model has a RMSE above four games and had an unsatisfactory R2, so another model was tested.

**4.2. Random Forest Model**

The next model tested was a random forest model. This first iteration used all predictors to create an estimate. After that model was created, the most important predictors were identified using a node purity score. A new model was then created using only the most important predictors, some of which were opponent three-point percentage, efficient field goal percentage (eFG%), seed, margin of victory (MOV), and true shooting percent (TS%). This model had a R2 of 0.518 which was similar to the linear model, however the model produced an RMSE of 3.31 on the test data which was an improvement. Predictions for the 2024 postseason are included in the table below.

**Table 2.** Comparison of Random Forest Model Predictions to 2024 Post Season

|  |  |  |  |
| --- | --- | --- | --- |
| ***Seed*** | ***Team*** | ***Wins*** | ***Win Predictions*** |
| 1 | OKC | 6 | 12.8 |
| 2 | DEN | 7 | 9.3 |
| 1 | BOS | 16 | 9.2 |
| 3 | MIN | 9 | 8.8 |
| 2 | NYK | 7 | 7.7 |
| 3 | MIL | 2 | 6.2 |
| 4 | CLE | 5 | 4.1 |
| 8 | NOLA | 0 | 4.1 |

The main strength of the random forest model was its low RMSE. However, the model had several shortcomings. It had a similar R2 to the linear model and incorrectly predicted the champion to be the Oklahoma City Thunder. As a result, another model was tried.

**4.3. Neural Network Model**

A neural network model was then tested. The model was created with the most important predictors from the random forest model. With this model, a very good RMSE was achieved on the training data, 2.59 games. However, when run on the test set, the RMSE increased to 3.57. There was an improvement in the R2 value, 0.70. So, the neural network model was able to explain about 20% more variance in the target variable than the other two models. The table below shows how the model performed on the 2024 post season.

**Table 3.** Comparison of Neural Network Model Predictions to 2024 Post Season

|  |  |  |  |
| --- | --- | --- | --- |
| ***Seed*** | ***Team*** | ***Wins*** | ***Win Predictions*** |
| 1 | OKC | 6 | 15.2 |
| 1 | BOS | 16 | 11.2 |
| 2 | DEN | 7 | 10.9 |
| 3 | MIL | 2 | 9.7 |
| 3 | MIN | 9 | 7.3 |
| 2 | NYK | 7 | 6.2 |
| 4 | LAC | 2 | 4.1 |
| 4 | CLE | 5 | 2.6 |

When tested on the 2024 playoffs, the model incorrectly predicts the NBA champion to be the Oklahoma City Thunder. This model does however boast the highest R2 and a RMSE less than the goal of four games. Another point to note is how strongly seeding influences predictions. Above are the eight teams with the highest projected wins; they are the top four seeds from both conferences.

**4.4. Ensemble Model**

Each model tested above had its own strengths and weaknesses. The MLR correctly predicted the 2024 NBA champion but also had the highest RMSE. The random forest model had the lowest test RMSE but an R2 of only 0.517, while the neural network achieved a higher R2 and acceptable test RMSE but relied heavily on the seed predictor. A model was needed to combine these models’ strengths.

In machine learning, stacking is an ensembling technique in which several base learners are trained, and a meta learner is trained to make a final prediction using the predictions of the base learners [2]. In our case, the MLR, random forest, and neural networks served as the base learners, and another linear model was used as the meta learner. The new stacked model achieved an R2 of 0.70 and a test RMSE of 3.55 games. Another point to note is that all three base learner predictions were deemed statistically significant in the summary output of the meta learner. The table below shows how the new model performed on the 2024 postseason.

**Table 4.** Comparison of Stacked Model Predictions to 2024 Post Season

|  |  |  |  |
| --- | --- | --- | --- |
| ***Seed*** | ***Team*** | ***Wins*** | ***Win Predictions*** |
| 1 | BOS | 16 | 12.9 |
| 2 | DEN | 7 | 10.7 |
| 1 | OKC | 6 | 10.0 |
| 3 | MIL | 2 | 9.7 |
| 3 | MIN | 9 | 9.2 |
| 5 | DAL | 13 | 7.7 |
| 4 | LAC | 2 | 6.5 |
| 6 | PHX | 0 | 6.1 |

This final model correctly predicted the 2024 NBA champion, achieved the highest R2 of any model, and had an acceptable RMSE of 3.55 on the test set. Not only that, it also closely followed the predictions of some sports betting agencies [5]. In the table below are the betting odds for the 2024 playoffs. Note that there is much crossover between the model’s predictions and the betting odds.

**Table 5.** Betting Odds for teams in the 2024 playoffs

|  |  |
| --- | --- |
| ***Team*** | ***Odds*** |
| BOS | +160 |
| DEN | +300 |
| LAC | +1300 |
| MIL | +1400 |
| PHI | +1400 |
| DAL | +1600 |
| OKC | +1600 |

Both the final model and this sports betting agency gave the Boston Celtics the best odds to win the championship and the Denver Nuggets the second-best odds. Another interesting point is that the runner up of the 2024 playoffs, the five-seeded Dallas Mavericks, were given the sixth-best odds to win the title in both my model and this sports betting organization.

5. Model limitations

Although the final model performed well, it still has its shortcomings. There are variables for which the model does not account. Injuries are one of these oversights. Neither this model nor any model can predict whether a team will stay healthy during the playoffs. Also, this model assumes that the regular season statistics are an accurate representation of the team. It cannot account for cases where a team’s star player missed significant time, leading to misleading regular season statistics.

Another limitation of this model is its inability to account for streakiness. Often, teams get hot during the playoffs and make an unexpected run. This is the human element of sports that machine learning models struggle to account for. Another oversight of the model is playoff experience. The created model only considers the regular season statistics to make its predictions; it does not consider whether a team has had success in past postseasons. Future research addressing these gaps could improve model performance.

**6. CONCLUSION**

This project has explored the concept of using regular season performance as a predictor of post season success in the NBA. Several models were developed and tested with this goal in mind, with the final model being a stacked ensemble that combined the strengths of a linear regression model, a random forest, and a neural network. While the ensemble model performed well in terms of both RMSE and R², it is important to remember that basketball and its players cannot be reduced to a machine learning model. We watch sports precisely because they are unpredictable.

That said, the model does have practical applications. A team’s front office could utilize a model like this to assess whether their team is capable of making a deep postseason run, potentially influencing how they approach a trade deadline or the offseason. Similarly, sports betting analysts could use the model’s outputs to validate or challenge betting lines. While no model will ever be perfect, this work demonstrates that data-driven analysis and machine learning can provide meaningful insights into postseason outcomes–insights that can be leveraged to make smarter decisions both on and off the court.

**ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to Professor Joe Faith for his guidance and sage advice throughout this project. His expertise played a pivotal role in the success of this undertaking.

I am also thankful to all my previous math teachers and professors who fostered my love and appreciation for math.

Additionally, I acknowledge my first basketball coaches, Ty and Becky Yager, who introduced me to the game I love.

Finally, I want to thank my parents, Reggie and Amy Weaver, who have shown me much love and support throughout all my life and academic career.

references

1. *Basketball statistics & history of every team & NBA and WNBA players | Basketball-Reference.com*. (n.d.). Basketball-Reference.com. https://www.basketball-reference.com/
2. Boehmke, B., & Greenwell, B. (2000, February 1). *Hands-On Machine Learning with R: Chapter 15: Stacked Models*. https://bradleyboehmke.github.io/HOML/stacking.html
3. Cox, T. (2022, April 18). *Updated 2022 championship odds for all 16 NBA playoff teams.* SportsBettingDime. https://www.sportsbettingdime.com/news/nba/updated-2022-championship-odds-all-16-playoff-teams/
4. Editors of Realhoopers. (n.d.). *How do the NBA playoffs work?* Realhoopers. https://realhoopers.com/how-nba-playoffs-work/
5. Taylor, C. (2024, April 20). *NBA playoffs 2024: The complete championship odds for every team*. The Rookie Wire, USA Today. https://therookiewire.usatoday.com/story/sports/nba/rookie/2024/04/20/nba-playoffs-2024-the-complete-championship-odds-for-every-team/77595762007/