Predicting NBA Players' Playoff Performance

Aaryan Jain, Aarav Shyamkumar, Amaya Malik, Siddakk Chatrath

Abstract

In this study, we developed a machine learning model to predict NBA players' playoff performances using regular season statistics and player age, drawing from data sources including the NBA's official statistics and an extensive dataset on player ages. Our linear regression model, which explained approximately 54% of the variance in playoff performance, was evaluated with a Mean Squared Error of around 23. The model's predictions, while moderately successful, suggested areas for potential refinement. Despite this, the results indicate a strong potential for machine learning in sports analytics to assist coaches and analysts in strategizing for playoffs, with further model modifications and the integration of additional variables to enhance prediction accuracy.

Introduction

In professional basketball, predicting playoff performance based on regular season statistics is a complex but an important part of sports analytics. Players showcase a wide range of skills and abilities throughout the regular season, but it's during the high-pressure playoff games where their performance may become different. Coaches and teams are often faced with the challenge of strategizing for the playoffs, making decisions that depend on understanding how a player's regular season performance might result in playoff success. Based on this understanding, the questions we decided to focus on are: "How can machine learning be utilized to analyze historical player statistics and predict their future playoff performance, taking into account variables such as regular season performance?" and "How does age impact player performance, and are older players valuable to a team?" With our project, we worked on answering these questions by making a machine learning model that uses historical data to predict players' performance in the playoffs. We came up with this approach by assuming that a player's regular season statistics – points scored, assists, rebounds, and other key metrics – indicate their potential playoff impact. Also, factors like a player's age are used, understanding that both experience and physical ability play important roles in high-pressure games. Our primary data sources for this analysis include the official NBA statistics website (https://www.nba.com/stats/leaders) and a dataset, all_seasons.csv, which provides information on players' ages. This project involves a number of steps starting from web scraping NBA statistics to processing and analyzing the data with advanced machine learning techniques. Our goal is to make a predictive model that offers insightful forecasts about playoff performances. We hope that these predictions will be valuable to coaches and analysts, helping them in making strategies and decisions that enhance team performance and adapt to the competitive dynamics of playoff basketball.

Data Description

The basis for our analysis is from available data online that shows a connection between lifestyle factors and chronic health outcomes. Research articles, such as the one we found in the Journal of the American Medical Association, have identified poor dietary habits, lack of physical activity, and insufficient sleep as primary contributors to the onset of chronic diseases. In professional sports, especially the NBA which we are looking at, athletes' performance can really be influenced by these lifestyle factors due to the demanding nature of their job. We chose to use the NBA stats website as it offers a unique opportunity to study these questions raised, as it provides a variety of performance data that can be publicly accessed by us and analyzed. We also used an additional dataset, 'all_seasons.csv', which contains historical data on NBA players' ages. This dataset was important in contextualizing performance data within the age demographic of the players, allowing for a more varied analysis of how age may impact performance over time. Our primary dataset was obtained through web scraping the NBA's official statistics website, which is known for its wide collection of player performance metrics. We wrote a diverse amount of functions, each designed to execute a specific task in the data retrieval and preprocessing pipeline:

1. **get_nba_data()**: This function was responsible for extracting statistics for each NBA player across regular and playoff seasons from the specified years (2012-2013 to 2020-2021). The data that we got from it included points per game, assists, rebounds, and other performance indicators.

2. **filter_nba_data()**: After we got the data, we used this function to filter it in order to isolate the statistics of players who were active in both regular and playoff seasons, thus making a uniform dataset for comparative analysis.

3. **get_player_age():** To introduce a demographic dimension to the analysis, players' ages were appended to the dataset by cross-referencing with 'all_seasons.csv', a historical player dataset.

4. **combine_df()**: The final step in our data preparation involved merging the regular and playoff season data into a detailed dataframe to facilitate integrated analysis.

The analytical phase comprised a variety of approaches to break down our data:

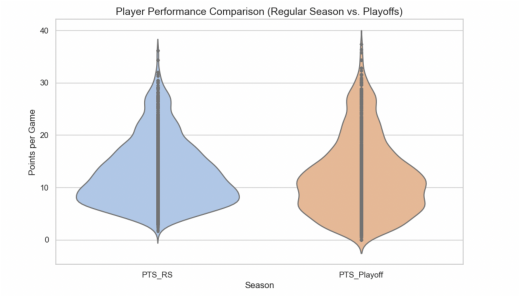
1. **Descriptive Statistical Analysis**: We started with a descriptive analysis to create a foundational understanding of the dataset, by calculating mean, median, and mode for different performance metrics.

2. **Comparative Analysis**: We then conducted a comparative analysis to identify gaps in performance between regular seasons and playoffs, which was important in understanding how players' performances varied across different competitive conditions.

3. **Regression Analysis**: To explore the relationship between age and performance, we made regression models, allowing us to predict performance metrics based on the age of the players.

4. **Longitudinal Study**: We also started a longitudinal study of individual players, such as Carmelo Anthony, to look at the progression of performance over time and correlate it with the players’ career arcs.

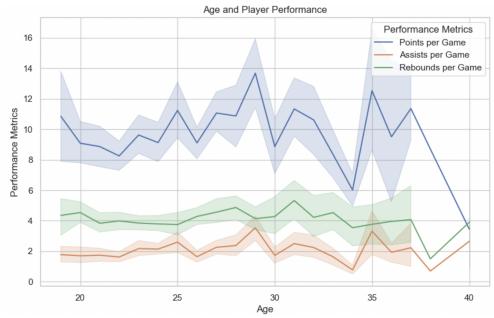
To visually show the findings from our analysis, we created 3 important visuals:



1. **Player Performance Comparison (Regular Season vs. Playoffs):** This graph shows a detailed visual comparison of NBA players' scoring performances across different season stages using a violin plot. This plot type is useful for showing the full distribution of a dataset, showing patterns that might be overlooked in other charts such as bar charts or line graphs. In this graph, we have two violins to represent points per game (PPG) for two distinct segments of the NBA season: the regular season (PTS_RS) and the playoffs (PTS_Playoff). The horizontal axis labels the season segment, while the vertical axis shows the PPG, offering a direct visual comparison of scoring rates.

   Each violin shape has multiple parts of information: Within each violin, a box plot is there, which shows the median score and the interquartile range of PPG. The median provides a quick reference for the central tendency of each distribution. The width corresponds to the number of players scoring at that particular PPG level. A wider section of the violin means a higher density of players at that score, while a narrower section indicates fewer players. The symmetry of the violin around the central axis reflects the nature of the distribution. A perfectly symmetrical violin suggests a relatively even spread of scores above and below the median. But on the other side, asymmetry can show skewness in the data, where a larger number of players score either somewhat higher or lower than the median. The tails represent the extreme parts of the distribution, showing the minimum and maximum PPG recorded. Long tails can indicate a significant range between the highest and lowest scorers.

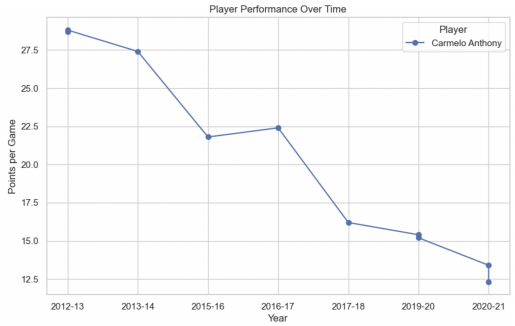
   The shape of each violin in the plot offers insights into scoring patterns, where a wider top on the playoff violin suggests increased scoring in high-pressure games, while a wider base on the regular season gives us a spread of lower scores for players. Variations in the smoothness of the distribution point to subgroups of scoring, with bumps signaling clusters of players who score similarly. The direct comparison between the two violins shows the differences in player performance across regular and playoff games, giving us information on consistency and the possible influence of game pressure on scoring.



2. **Age and Player Performance:** This graph shows the relationship between age and relevant basketball performance metrics through a multi-line chart, where each line goes through the average number of points, assists, and rebounds per game that NBA players achieve at various ages, offering a visualization of an athlete's performance life cycle. The lines are color-coded—blue for points, orange for assists, and green for rebounds—each providing information of how the different parts of a player's game might progress and regress as they become older.

   The shaded areas covering the lines are confidence intervals, which are important in understanding the data as they tell us the range within which we can be certain the true average performance metrics lie, considering the sample of players at each age. The narrower the shaded region, the higher our confidence in the precision of the average performance estimate for that age. These intervals are important when considering the variance in individual player development and aging, suggesting that while trends can be observed, there's some individual variability.

   The graph's peaks can tell us an athlete's prime years, where experience and physical ability intersect to give peak performance, whereas troughs could possibly point to transitional phases in a player's career, such as entering the league or approaching retirement. An upward trend in any metric may be because of the development of skills or roles players grow into as they gain NBA experience, while a downward slope could show the decline due to aging or the changing nature of a player's role on their team. Also, cross-metric comparisons reveal the balance between different elements of the game; for example, while scoring might decline with age, a player's assists may increase, possibly indicating a shift from a scoring-focused to a more teamplaying role. The intersection points where lines meet or cross over are informative to us, revealing ages where players might change their style of play, showing different aspects of their game.



3. **Player Performance Over Time:** This graph provides a longitudinal analysis of Carmelo Anthony's scoring skill by charting his points per game (PPG) across several NBA seasons, starting from 2012-13 and ending in the 2020-21 season. While the graph specifically looks at Carmelo Anthony's performance, the function we used to make this graph can be used to make similar data visualizations for any NBA player. This line graph serves not only as a statistical record but also to describe Anthony's career, showing how his role as a scorer has changed over time, which can be seen in the changes in PPG. The initial increase followed by a decline shows a career arc common to many professional athletes, where peak performance is achieved by a period of adaptation to changing physical capabilities and team contexts.

   Key moments, such as significant changes in team composition, leadership shifts, injuries, or personal developments, are shown by annotations, providing a look into the shifts in the graph's trajectory. For example, a sharp decline or increase could correlate with a trade to a new team, a change in team strategy, or recovery from injury, each affecting his average scoring output. The graph's ability to condense years of performance into a single visual timeline allows for a look at career stages, from his prime scoring years to the more later part of his career.

   Also, this focused analysis on a single player's scoring trend over nearly a decade offers a look into the lastingness of athletic performance, showing the impact of age and external factors on a player's performance. It also shows the value of changing and adapting playing style and expectations when a player's performance begins to decline, as well as the potential for a player to play better in a team over time.

Our extensive and detailed analysis shows diverse patterns of player performance. While some players show an improvement in their game during playoffs, others show a decline, suggesting that high-pressure situations might influence player performance. The age analysis reveals an expected decline in performance metrics with increasing age, stressing the physical demands of basketball. Individual player trajectories give us information about career progression, which is influenced by a combination of factors including personal, team-related, and possibly lifestyle-related factors.

What we concluded from our graphs and analysis is that while performance in professional athletes, such as NBA players, could be correlated with age, the relationship is varied and may also be impacted by psychological factors, team dynamics, and physical health, which are all combined with lifestyle choices.

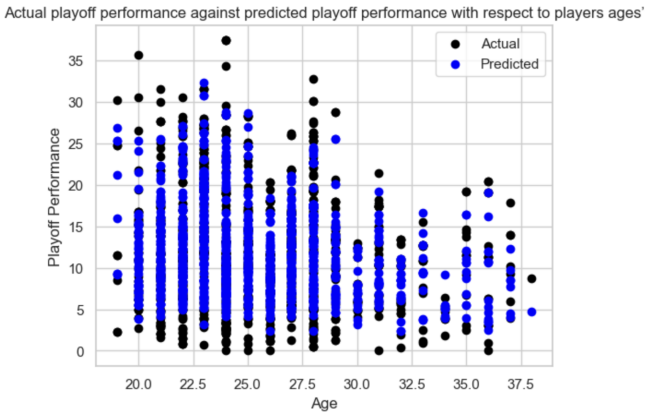
Method

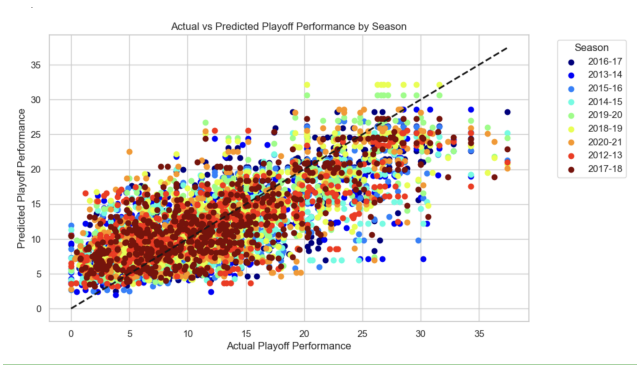
Linear Regression Model - The ML method we used is a simple linear regression model. After we collect and clean the data on every basketball player in the NBA, we could use feature selection as a way to give weightage for certain factors. For example, stats from their regular season and playoffs can be different or even their current age. Once the key features are identified, then influences on the playoff prediction model can be made. We'd train the model on historical data, where regular season and playoff stats are used to predict playoff performance as the target variable. From this, many assumptions can be made. First one can be that a linear relationship exists between the selected features and playoff performance. The features could be the key features we chose earlier, like regular season or playoff stats, or age. Also, this assumption might not always hold, and more complex relationships may exist. Another assumption is to ensure the independence of observations, meaning each player's data point is independent of others. One important part that we have to consider is being cautious of overfitting. We must avoid fitting the model too closely to the training data, as this may lead to poor generalization on new, unseen data. To assess the model's performance and prevent overfitting, cross-validation techniques we used k-fold cross validation. This involved dividing the dataset into k subsets, training the model on k-1 folds, and evaluating its performance on the remaining fold. Additionally, the model was tested on a separate test set to estimate its predictive accuracy for future playoff performances, providing an evaluation of how general our model can be. Lastly, we assess the model's performance using the right metrics like the Mean Square Error, and the R-Squared.

Linear Regression stands out as an ideal choice for predicting player playoff performance based on its simplicity, transparency, and interpretability. The coefficients that we got from the model directly reveal the influence of each factor, helping us with decision-making processes. The flexibility that it offers allows for the incorporation of various features, including regular season statistics, playoff metrics, and age, allowing for a detailed analysis of factors influencing player performance. While the model assumes a linear relationship between features and playoff performance, its design supports situations where the relationship may be more complex. To make it stronger, our implementation involves cross-validation techniques and provides the option for Polynomial Regression to also capture nonlinear patterns. Another important part is the practical efficiency of Linear Regression, combined with its computational efficiency, which gives a streamlined yet effective solution for predicting playoff performance without using other complex computational resources.

Results

Our machine learning model, based on linear regression, aimed to predict NBA players' playoff performances using regular season points and age as key features. The model was trained and tested on a dataset compiled from official NBA statistics and a kaggle file which was a dataset that held the player's ages.





The first graph is a scatter plot that shows actual playoff performance (black dots) against predicted playoff performance (blue dots) with respect to players' ages. The distribution of points suggests that for most age groups, there is a variance in performance prediction, with the model tending to underpredict for higher values and slightly overpredict for lower values. Notably, there's a concentration of predictions for players around the age of 25-30, which might indicate this is a common age range for players in the dataset. The spread of actual performance also appears wider than that of the predicted, hinting at the model's conservative nature in estimating performance.

The second graph shows the actual vs predicted playoff performance after cross-validation. The diagonal dashed line represents the line of perfect prediction where the actual performance would match the predicted performance. The points are mainly below this line, especially for higher performance values, indicating a tendency of the model to underpredict in cases of higher player performance. But besides this, the cluster of data points around the lower end of the spectrum shows that the model has decent predictive power for players with lower playoff performance metrics.

Discussion

Our linear regression model, applied to predict NBA players' playoff performance, exhibited a moderate level of success. By utilizing regular season statistics and player age as key predictors, the model explained approximately 54% of the variance in playoff performances. This finding is significant in the context of sports analytics, providing a foundation for predicting player performance in high-pressure playoff situations. However, the unpredictable nature of sports, influenced by numerous on-the-spot decisions, psychological factors, and team dynamics, means that these results should be used in conjunction with expert insights and real-time analysis for effective application.

The study initially aimed to determine the efficacy of machine learning in predicting playoff performances. Our results, with a Mean Squared Error of 23.33 and an R-squared value of 0.54, suggest a reasonable level of accuracy. The model's capacity to explain over half of the variance in data points towards its effectiveness in capturing trends and patterns in player performance. This indicates a partial solution to our question, highlighting both the potential and limitations of using statistical models in sports performance prediction.

While the results are promising, we should not fully accept them due to limitations of linear regression in capturing the complex, non-linear dynamics of sports performance. Despite the predictive power, our model simplifies the complex nature of player performance greatly. The MSE values, hovering around 23 in both standard and cross-validated models, point to a significant average difference between predicted and actual performances, suggesting room for improvement in model accuracy.

To increase the strength of the model’s accuracy, we incorporated K-fold cross-validation with 5 splits, achieving a slightly improved cross-validated MSE of 23.97 and an R-squared of 0.55. This improvement in the model's consistency across different data improves its generalizability and reliability.

Recommended Actions and Confidence Level
Considering these insights, we recommend:

- Exploring more complex, non-linear modeling techniques to better represent the intricate nature of sports performance.
- Expanding the dataset to include additional variables, such as psychological factors and in-game strategies, for a more accurate analysis.
- Continuously updating the model with new data to maintain its relevance and accuracy.

These recommendations are made with moderate confidence, recognizing the unpredictable elements of sports but also the growing potential of advanced analytics in enhancing performance predictions.

Unanticipated Questions and Future Directions
Our study raised several new questions, especially regarding the impact of external factors on player performance. Future research should focus on:

1. Incorporating complex algorithms to handle non-linear relationships and player performance dynamics.
2. Including comprehensive studies involving psychological assessments and team dynamics.
3. Conducting longitudinal analyses to track performance evolution over players' careers.

Overall, our study shows a significant step in using machine learning for predicting NBA playoff performances. While our current model offers valuable insights, the complexity of sports performance brings the need of deeper, more detailed analytical approaches. Future analyses, combined with diverse data and advanced modeling techniques, can potentially enhance the accuracy and practical utility of sports analytics, especially in the context of playoff performance prediction.

Citations

1. Source: https://stats.nba.com/stats/leagueLeaders?LeagueID=00&PerMode=PerGame&Scope=S&Season=2012-13&SeasonType=Regular%20Season&StatCategory=PTS
   By: NBA or nba.com
2. Source: https://www.kaggle.com/datasets/justinas/nba-players-data (Kaggle Source)
   By: Justinas Cirtautas
   File: all_seasons.csv