Dean&Francis

ISSN 2959-6157

NBA Player Data Statistics and Predictions Based on Spark

Yuchen Ye

Abstract:

Entering the 21st century, the sports field has begun to rely on advanced data to establish fine-grained athlete evaluation metrics, such as the impact of an athlete's presence on the team and performance in critical moments. These metrics are included in statistics and analysis. This experiment analyzes player datasets from NBA games, processing and statistically analyzing data indicators such as player rankings, changes brought by player age and career length, and changes in the proportion of three-point shots. The rating score proposed in this study is used as a measurement indicator to predict Luka Doncic's performance since entering the league. The model's performance is evaluated by calculating the root mean square error between the predicted and actual data. Analyzing and predicting player game data is crucial for evaluating player performance, formulating tactics, and management strategies. It not only helps management and coaching teams make more informed draft and trade decisions but also reveals similarities and differences among players through techniques such as cluster analysis, guiding personalized training and tactics.

Keywords: NBA, Spark, data processing, data analysis, prediction model

1. Introduction

With the rapid development of big data technology, sports data analysis has become an important means to enhance athlete performance and formulate game strategies. Basketball, as a data-rich sport, contains a wealth of valuable information in its game data. However, traditional data processing methods struggle to meet the computational demands of massive datasets. Spark, as an efficient distributed computing framework, provides powerful data processing capabilities, enabling the processing and analysis of large-scale datasets in a short time.

This study aims to explore big data processing and analysis methods based on Spark, conducting a comprehensive evaluation of NBA player performance. We will detail the data preprocessing process, data analysis and processing, feature extraction methods, and the construction process of future prediction models based on machine learning. By using scientific statistical methods, we can provide data support for basketball players' training and games, helping coaches and players make more informed decisions.

2. Experimental Steps and Process Display

2.1 Player Stats Preprocess

The dataset comes from the NBA official website (https://www.nba.com/stats/), covering player statistics from all regular seasons from the 1996-1997 season to the 2023-2024 season. First, it is necessary to remove empty rows and useless header rows from the dataset and handle outliers in the data. The asterisks in player names (indicating All-Star selections) should be removed, and missing values (represented by empty fields between two commas) should be replaced with 0.

Next, data integration is performed, combining player statistics from each season from 1996 to 2023, and adding year information to each row of data. Each row of data in the CSV file is split into multiple fields to facilitate subsequent analysis and processing. Finally, the processed data is saved as CSV files, partitioned by year, to facilitate year-by-year analysis.

2.2 Basic Stats Analyze

First, players who have played more than 30 games in each season are selected to avoid extreme data losing its reference value. The article mainly extracts and saves the following five types of statistical data: Points (PTS), Assists (AST), Rebounds (TRB), Steals (STL), and Blocks (BLK). These statistics are categorized by player name, year, and team, and sorted in descending order. The top ten results for each indicator are visualized as follows:

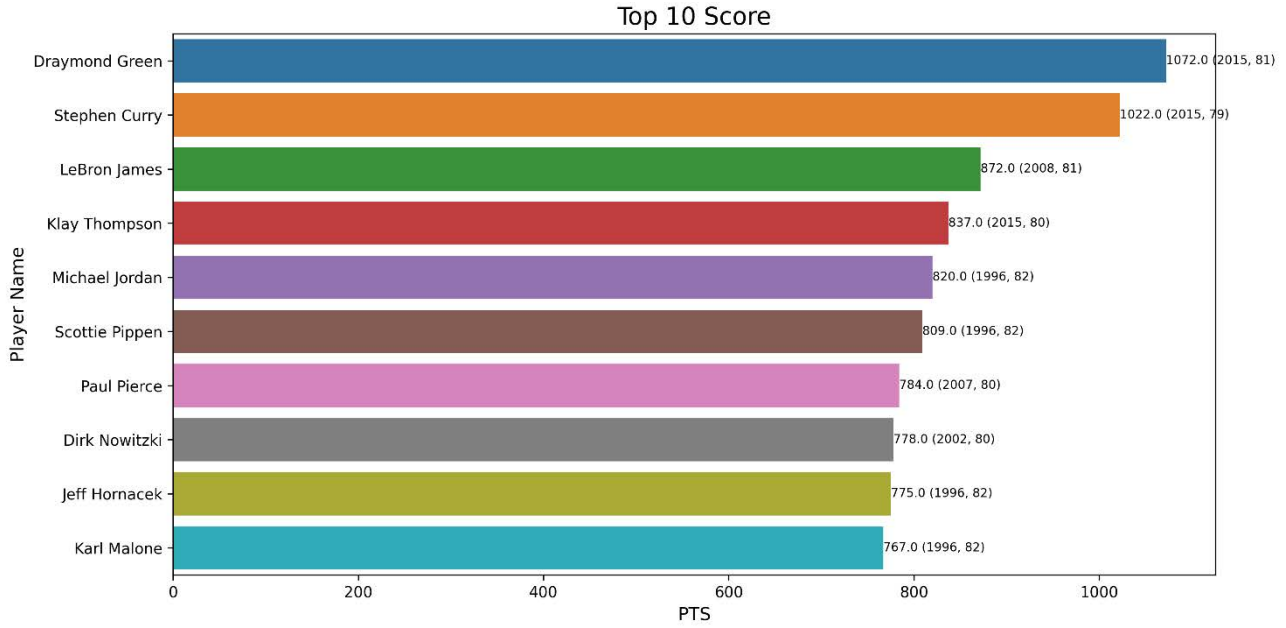

Figure 1 Shows the names of the top ten players in scoring and their respective scores.

ISSN 2959-6157

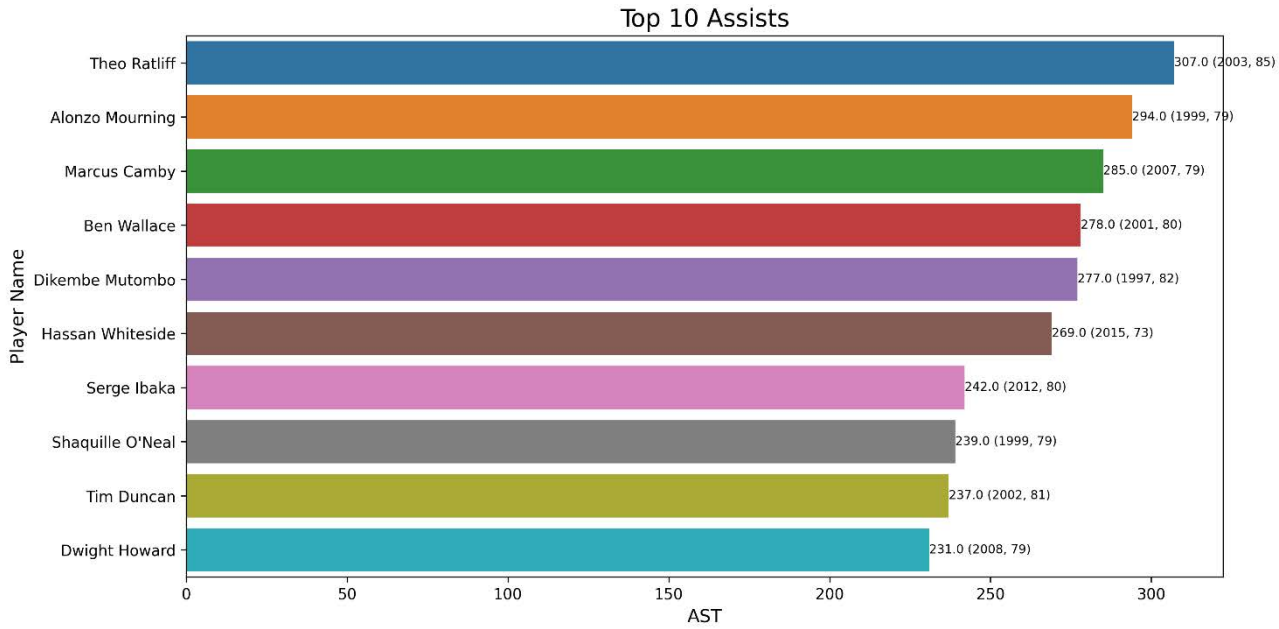

Figure 2 Shows the names of the top ten players in assists and their respective assist numbers.

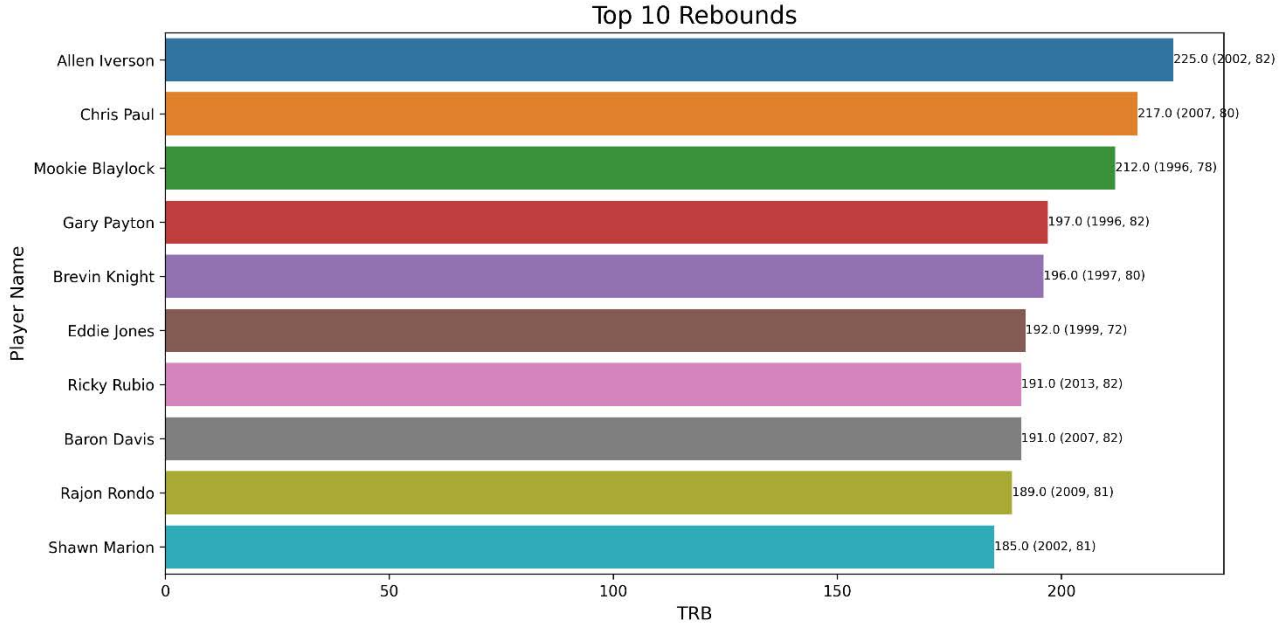

Figure 3 Shows the names of the top ten players in rebounds and their respective rebounds numbers.

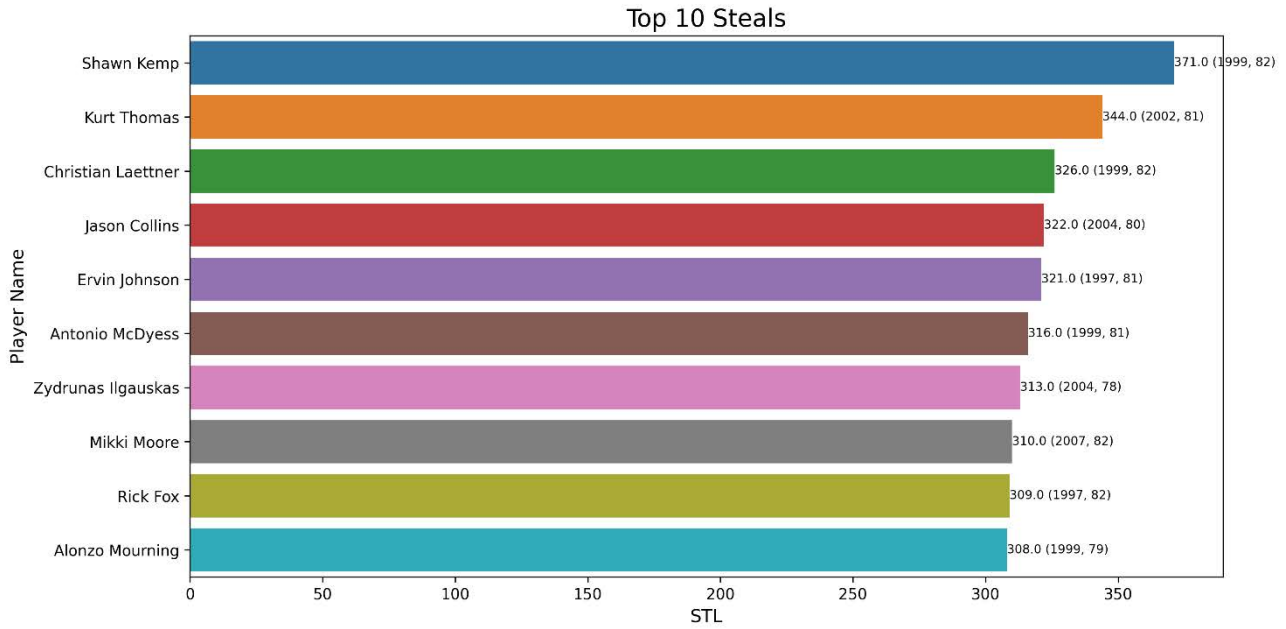

Figure 4 Shows the names of the top ten players in steals and their respective steals numbers.

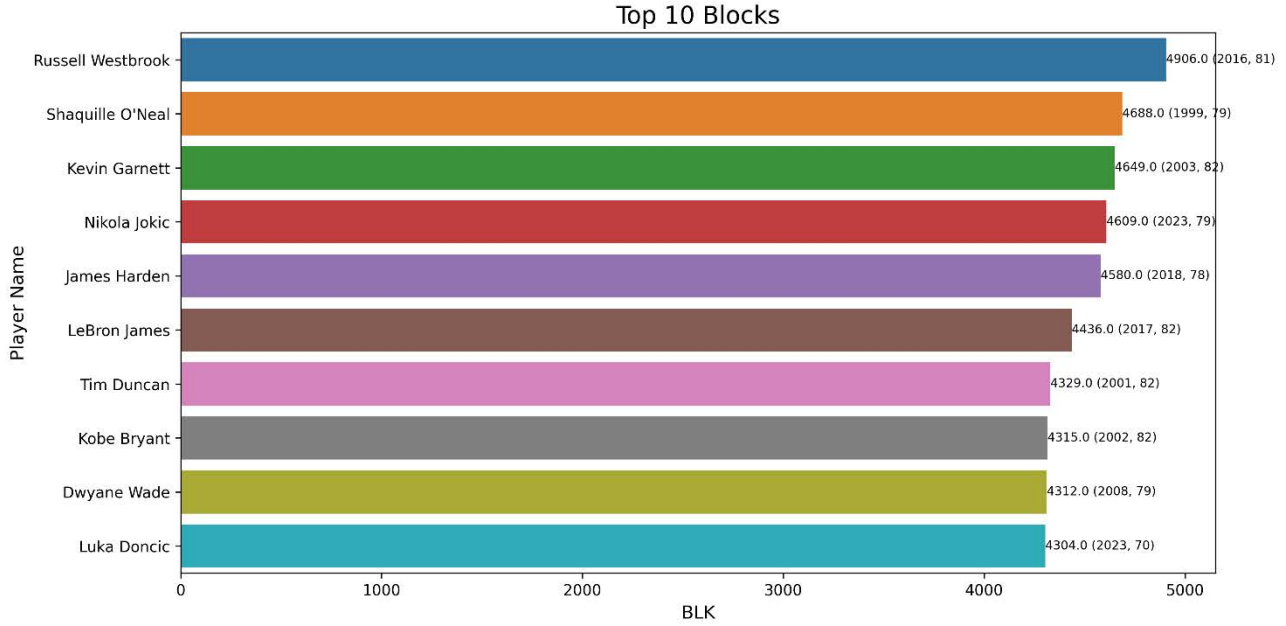

Figure 5 Shows the names of the top ten players in blocks and their respective blocks numbers.

2.3 Age Exp Trend Analyze

Through a detailed trend analysis of the relationship between NBA players'age and experience with their game performance (points, rebounds, assists, etc.), this section reveals the impact of these factors on player performance at different age and experience levels. The calculation of player experience is based on the player's age to estimate their years of experience in the league. The experience value is calculated as: Experience = Current Year - Player's Age + 1.

The trend analysis includes two parts: age trend analysis and experience trend analysis.

Age Trend Analysis: Statistics on player performance based on age, analyzing performance differences across age groups. This includes the average, standard deviation, maximum, and minimum values of statistical data such as points (PTS), rebounds (TRB), and assists (AST).

Experience Trend Analysis: Statistics on player perfor-

ISSN 2959-6157

mance based on years of experience, analyzing performance differences across different experience groups. This includes the average, standard deviation, maximum, and minimum values of statistical data such as points, rebounds, and assists.

The visualized line charts are shown below, the vertical thin lines, known as error bars, represent the standard deviation of the data. The longer the error bars, the greater the variation in player performance within the age/experience group. The shaded area indicates the range of the data, which is the interval between the minimum and maximum values.

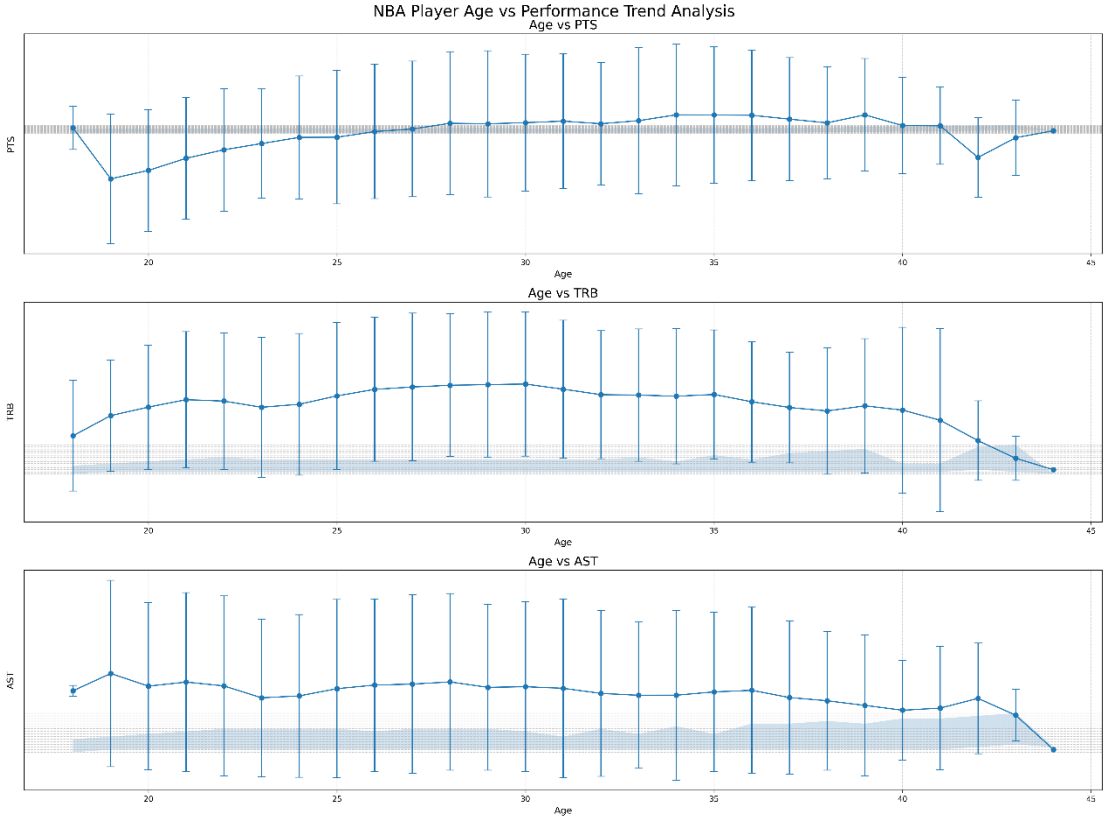

Figure 6 Shows the relationship and trend between player age and performance.

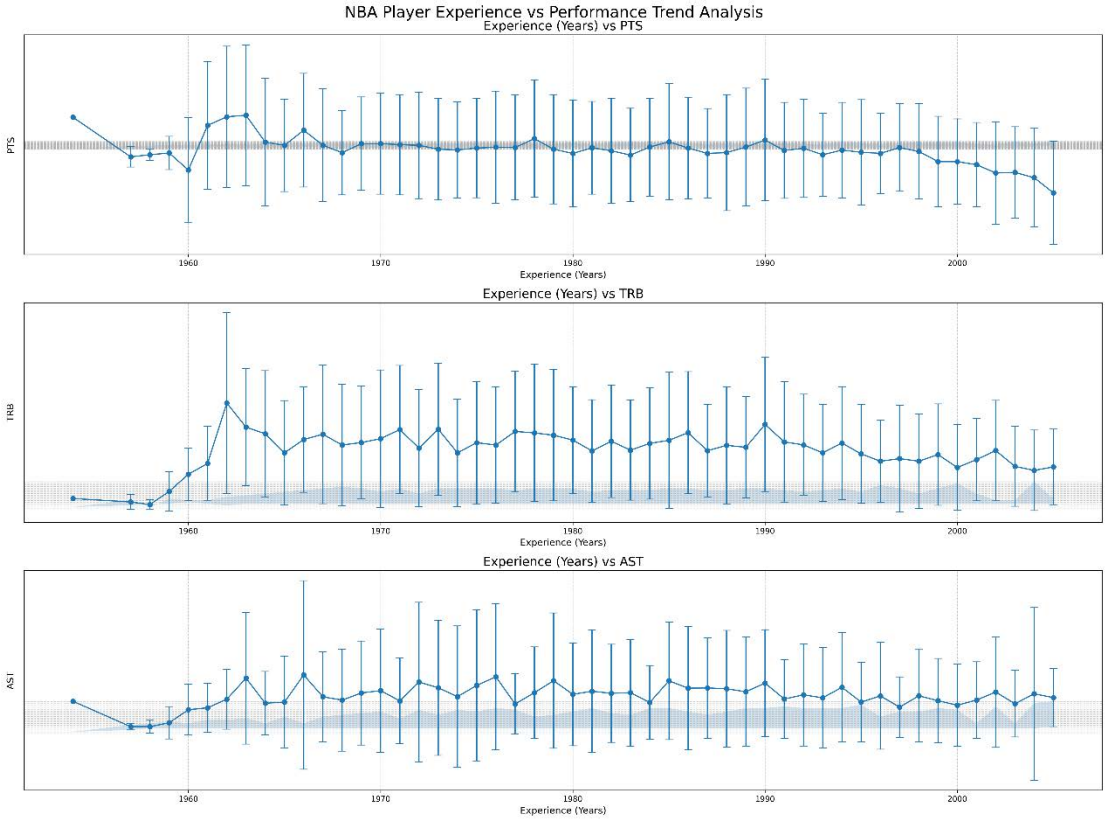

Figure 7 Shows the relationship and trend between player experience and performance.

2.4 Team Change Analyze

First, map the data to a Team class that includes player names and teams. Then, group the data by player and count the number of different teams each player has played for during their career, which represents the number of team changes. Next, group the players by the number of team changes and count the number of players for each number of team changes. The bar chart is arranged in ascending order based on the number of team change.

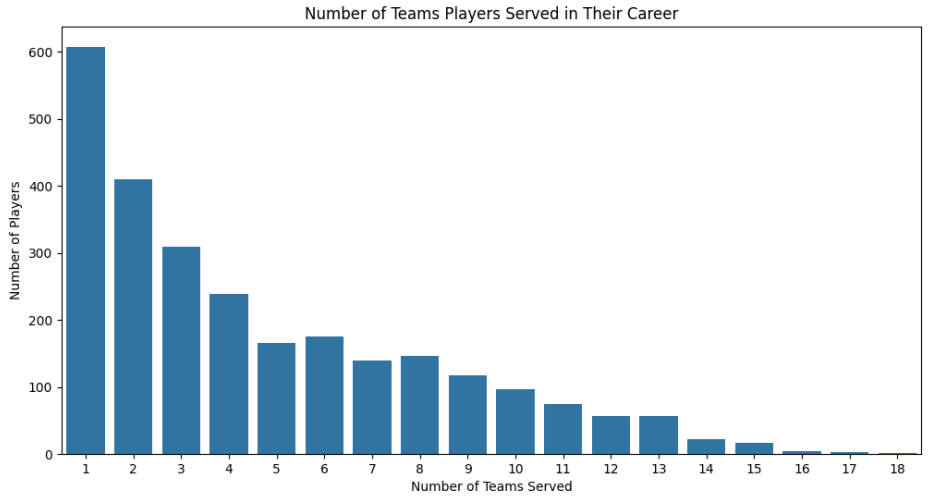

Figure 8 Shows the number of teams players served in their career.

ISSN 2959-6157

2.5 Three Point Analyze

Three-point shooting is a crucial scoring method in basketball. This analysis focuses on examining the number of three-point goals made on average players, the number of three-point attempts, and the three-point field goal percentage over time. First, player data with more than 30 games played is selected to ensure the quality and relevance of the analysis. The filtered data is mapped to a custom data structure (Three_P case class), including the year, number of successful three-point shots, number of three-point attempts, and three-point field goal percentage. Then, the three-point data for each year is aggregated to calculate the total number of three-point goals made on average players and attempts, as well as the average three-point field goal percentage.

From Figures 9 and 10 are shown below, it can be seen that over time, the tactical development of NBA games and the athletic level of players have been improving. Three-point shooting has become increasingly important in games, not only reflected in the higher proportion of points scored from three-pointers but also in the increasing number of three-point attempts and shooting percentages by players.

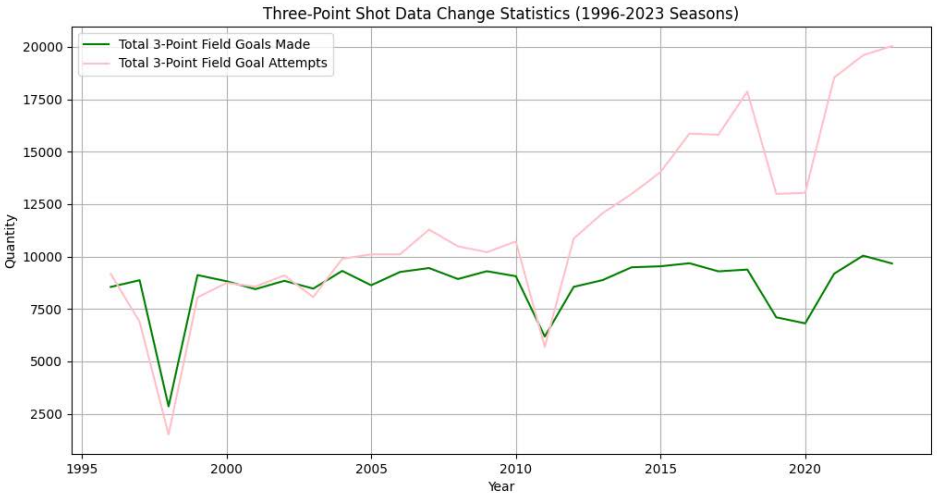

Figure 9 Shows the trend and change of three-point data.

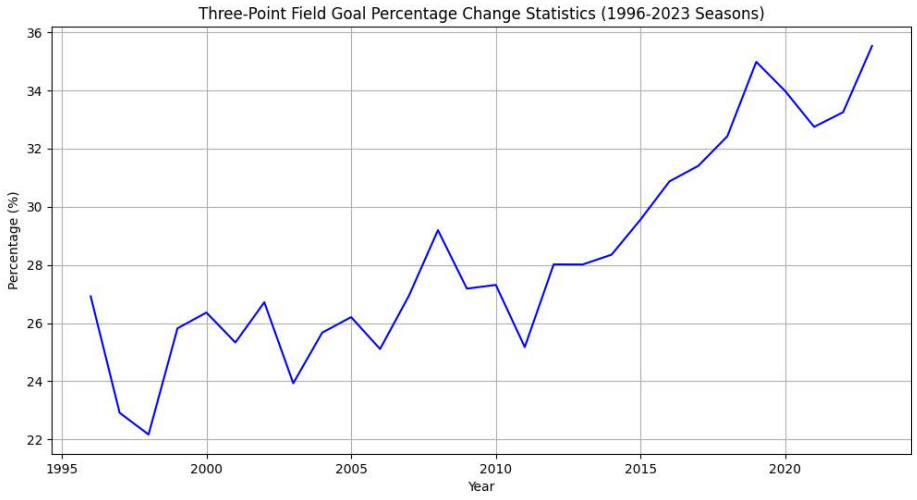

Figure 10 Shows the trend and change of three-point field goal percentage.

2.6 Predictions for Luka Doncic

The rating calculation system aims to fairly and objectively evaluate a player's annual performance from the perspective of game data. The rating is calculated by computing the average and standard deviation of various statistical data each year and normalizing them. The formula for calculating the rating is as follows:

$$rating = \frac{X - \mu}{\sigma}$$

Where (X) is a specific statistical data point for the player, (μ) is the mean value of that data point, and (σ) is the standard deviation of that data point.

Normalization of the score:

$$nR = \frac{r - r_{min}}{r_{max} - r_{min}}$$

Where (r) is the player's rating score, and (r_{min}) and (r_{max}) are the minimum and maximum rating scores for that data point in that year.

Total Rating Score:

$$rTOT = rFG + rFT + r3P + rTRB + rAST + rSTL + rBLK + rTOV + rPTS$$

where each (r) score represents the rating score for different statistical data.

Normalized Total Rating Score:

$$nTOT = nFG + nFT + n3P + nTRB + nAST + nSTL + nBLK + nTOV + nPTS$$

where each (nR) score represents the normalized rating score for different statistical data.

Calculating Player's Career Length: Calculate each player's career length (i.e., the player's age minus the player's minimum age).

Feature Engineering and Model Training: Use the normalized (R) scores as features to construct feature vectors. Train a random forest regression model with the target variable being the total rating score ($rTOT$) to predict player performance.

Model Evaluation and Prediction: Evaluate the model performance on the test set by calculating the root mean square error (RMSE).

Predicting Future Performance for Specific Player ("Luka Doncic"):

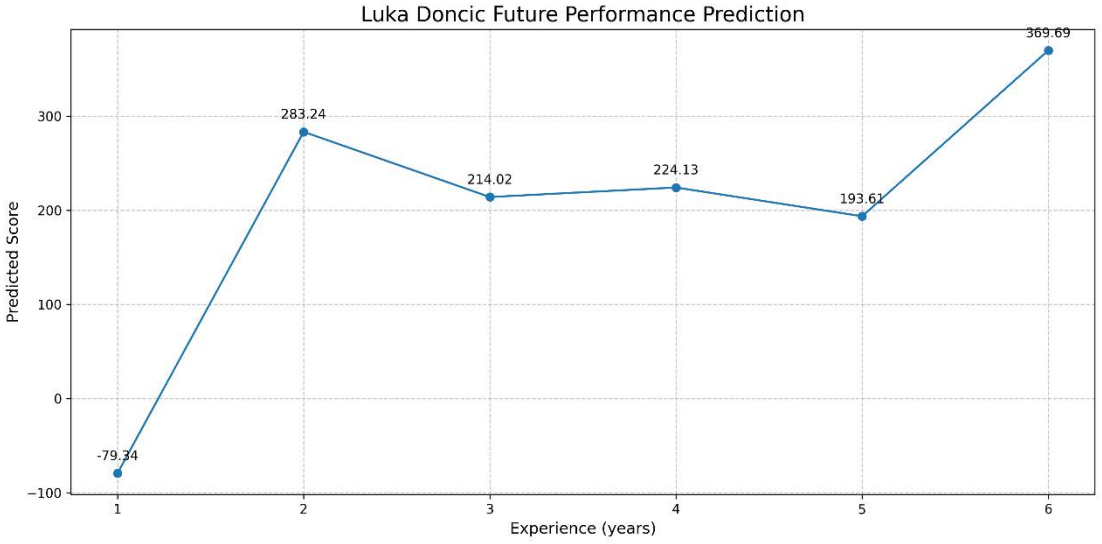

Figure 11 Shows the prediction trend of Luka Doncic's future performance.

Model Performance Evaluation:

The RMSE (Root Mean Square Error) is the error of the model trained on data from 1996-2022 when tested on the test set. The lower the RMSE, the better the model's predictive performance.

RMSE (1996-2022): 54.12

RMSE (2023): 41.36

The RMSE value of the model trained on 2023 data is lower than that of the model trained on 1996-2022 data, indicating that the 2023 model performs better in predictions.

ISSN 2959-6157

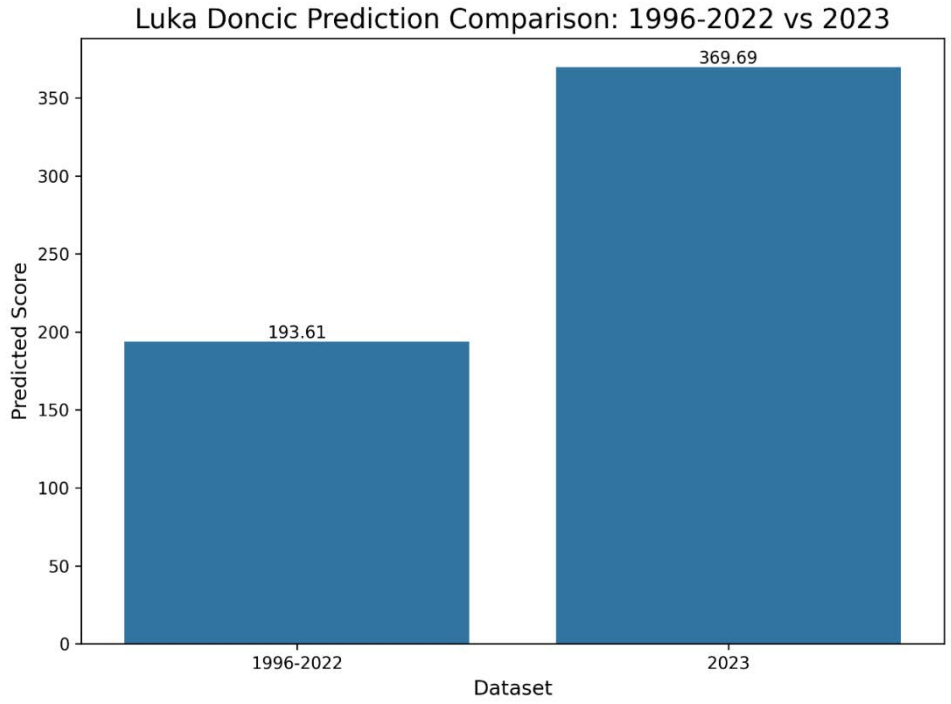

Figure 12 Shows the comparison of using two different training models when predicting Luka Doncic's future performance.

From Figure 12, the comparison of predicted and actual data for Luka Doncic at different ages (e.g., 22 to 24 years old) is shown. The "features" column displays the input feature vectors, while the "prediction" column shows the model's predicted values for PTS (points). In the 1996-2022 model, the predicted values for Luka Doncic's points vary significantly, possibly because the model was not specifically optimized for 2023. Due to the long span of training data, the model's predictions may be influenced by historical data, leading to less accurate predictions for the latest data.

In contrast, the 2023 model's prediction result for Luka Doncic's points in 2023 is 369.69. This model was specifically trained on 2023 data, making the prediction more closely aligned with the actual performance in 2023.

Overall, the 2023 model is more accurate in predicting Luka Doncic's performance compared to the 1996-2022 model. This indicates that training data closer to the prediction year generally yields better predictive performance.

3. Conclusion

Through in-depth analysis and model prediction of NBA player data from 1996 to 2023, we have revealed various factors influencing player performance and provided valuable references for future player performance predictions. During this period, the NBA has transitioned from traditional basketball to modern basketball, with the increasing importance of three-point shooting and significant changes in player training and game strategies.

Our research first ensured data integrity and consistency through data cleaning and integration, laying a solid foundation for subsequent analysis. Basic statistical analysis extracted key performance data of players, providing a basic overview of player performance. Age and experience trend analysis revealed trends in player performance with increasing age and experience, while team change analysis showed the impact of team changes on player performance.

The analysis of three-point shooting data, in particular, reflects the importance of three-point shooting in modern basketball. Players with more game appearances showed significant performance in three-point shooting percentage and attempts, providing important references for team tactics and player training. Detailed data analysis of Luka Doncic demonstrated his performance and development trajectory across different seasons, further enriching our research findings.

The random forest regression model used to predict player performance showed that the model trained on 2023 data was more accurate in predicting Luka Doncic's performance compared to the model trained on 1996-2022 data. This indicates that the closer the training data is to the prediction year, the better the predictive performance.

Overall, this study not only reveals various factors in-

fluencing NBA player performance but also reflects the development and changes in basketball over the past decades. Future research can further optimize the model by incorporating more external factors, such as injury records and training intensity, to improve prediction accuracy and practicality. These findings provide valuable insights for team management and player development and offer strong support for the future development of basketball.

References

[1] N Zhu, Q Dai.Basketball Data Analysis Based on Spark Framework and K-means Algorithm.The 2021 International Conference on Machine Learning and Big Data Analytics,2022•Springer

[2] Y Zhao, MF Ramos, B Li.Integrated.framework to integrate Spark-based big data analytics and for health monitoring and recommendation in sports using XGBoost algorithm.Soft Computing,2024•Springer

[3] LR Nair, SD Shetty, SD Shetty.Applying spark based machine learning model on streaming big data for health status prediction.Computers & Electrical Engineering,2018 - Elsevier

[4] P Sewal, H Singh.A machine learning approach for predicting execution statistics of spark application.on Parallel, Distributed and Grid Computing,2022 - ieeexplore.ieee.org

[5] K Wang, MMH Khan.Performance prediction for apache spark platform.on Cyberspace Safety and Security,2015•ieeexplore.ieee.org

[6] J Reece, SY Hong.Big data analytics for smart sports using apache spark.Issues In Information Systems,2021•iacis.org