Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

NBA Player Comprehensive Score Prediction based on Linear Regression and Random Forest

Junliang Lyu [1], Ouze Wang [2], Yinghao Zhang [3, *]

[1] International Business, Jinan University, Guangzhou, China

[2] ShengHua ZiZhu Academy, Shanghai, China

[3] Suzhou Dulwich High School, Suzhou, China

* Corresponding Author Email: alex.zhang25@stu.dulwich.org

Abstract. This paper introduces machine learning models for constructing a comprehensive quality prediction model to forecast National Basketball Association (NBA) players' specific scores. The objective is to facilitate the analysis, guidance, and evaluation of players' relevant value. Initially, data processing involves renaming the dataset and dividing it into an 80% training set and a 20% dataset through preprocessing, simultaneously addressing missing row. Subsequent steps include visualizing the data and conducting correlation analysis by group, producing a correlation heatmap to mitigate multicollinearity issues. Based on the visualization chart's summary, a tentative ability map of players across different positions is delineated, covering rebounds, assists, and other aspects. Employing random forest and linear regression methods, NBA player data is utilized to train the model, followed by comparison of different models' performances and analysis of their respective strengths. Histograms and linear graphs for the linear regression and random forest models are derived, with random forest exhibiting superior fitting to the data, indicating more accurate predictions compared to linear regression. For future projects, the aim is to employ a diverse range of models for comprehensive data analysis and utilize various evaluation methods for detailed assessments of the models.

Keywords: Machine learning; Random Forest; Linear regression.

1. Introduction

Given the recent trend in Data Science (DS) and Sports Analytics, an opportunity has arisen for utilizing machine learning (ML) techniques in sports [1]. This paper reviews the scores and chances each player gain in National Basketball Association (NBA) in 2023 season. The purpose of this paper is to use machine learning and data analyzing to predict the score a player could get in one single game. By predicting the scoring performances of each player and summing them up, it can be determined if predicting individual performances can be used an accurate win-loss classifier [2]. Furthermore, it serves as a crucial performance metric that allows coaches, analysts, and fans to assess a player's scoring ability and overall offensive contribution to the team. We encourage managers and coaches of sports teams to choose appropriate methods according to their aims. Future research should take into consideration the use of models with random effects on players' characteristics [3].

With the advancement of Machine Learning methods and the increasing demand of predicting players scores, Machine Learning has become a popular method to predict NBA games. Matthew Houde has used machine learning methods to improve the traditional NBA game prediction and makes the accuracy up to 65.1% [4]. Kuan-Chieh Wang uses machine learning, especially variants of neural networks, to achieve automated detection and data analysis of competitions [5]. Albrecht Zimmermann has refined the information as team, rebounds and assists, and through regression model analysis in machine learning, it concluded that the correlation between game results and player injuries and home advantage is relatively high, while the correlation with basic technical data of the game is relatively low [6]. And in In Estimate the Ability of NBA Player, Paul Fear head let us know the specific evaluation criteria for evaluating the abilities of NBA players He divides players' abilities into on-court and off-court, offensive and defensive aspects, and based on comprehensive data

analysis, he concludes that traditional NBA evaluations overemphasize players' offensive capabilities while overlooking their defensive abilities [7]. In addition, there is a noticeable point that, when there is a large amount of National Collegiate Athletic Association of Basketball (NCAAB) data on the network, Albrecht Zimmermann proposes that NBA and NCAAB data need to be classified and analyzed, otherwise it may lead to inaccurate data predictions [8]. Bryan Cheng's analysis report is based on research in the field of sports prediction, using various data to make more accurate predictions of NBA game win rates, which is helpful for us to analyze player value [9].

The main purpose of this study is to construct a comprehensive quality evaluation model by introducing machine learning models, in order to predict the scores of individual NBA players. Specifically, firstly, machine learning models including linear regression and random forest are introduced to establish benchmark models. The characteristics of NBA players are collected to train the model. Secondly, analyze the advantages of different models by comparing their predictive performance. Finally, the models' final prediction results can be effectively judged as efficient predictions, with a sufficiently high R2 score. We believe this result can help analyze, coach, and the entire association evaluate the value or ability of players, thereby optimizing their tactics and decision-making.

2. Methodology

2.1. Dataset Description and Preprocessing

The dataset used in this study, called 2023_nba_player_stats, is sourced from Kaggle [10]. The DataFrame contains basketball performance data for players. The columns of the DataFrame have original names that may not be user-friendly or self-explanatory. Therefore, we renamE the columns to more meaningful and understandable names. One of the columns is named "Position" and represents the position of each player. However, some of the rows in the "Position" column have missing values (NaN). We use the position "SG" (shooting guard) to fill in the missing values for all players with unknown positions. The dataset is preprocessed by splitting it into training set and test set. The independent variables (features) are stored in X and the dependent variable (target) is stored in y. The training set contains 80% of the data and the test set contains the remaining 20%. The training set contains 80% of the data and the test set contains the remaining 20%. The randomization state is set to 42 to ensure repeatability.

2.2. Proposed Approach

Our main idea is using machine learning to achieve prediction of NBA players points. First, we read dataset and analysis it. We changed a lot of names to be more meaningful and understandable. After that we did data visualization. This step helped us to see the correlations between all the features as we can see in the dataset. The final correlation map is the best way to choose features. After we selected the features, the main object is training the models: linear regression and random forest tree. linear regression model and random forest regression model are created and tested using different test sizes and random states to find the optimal configuration. The pipeline is shown in the Figure 1.

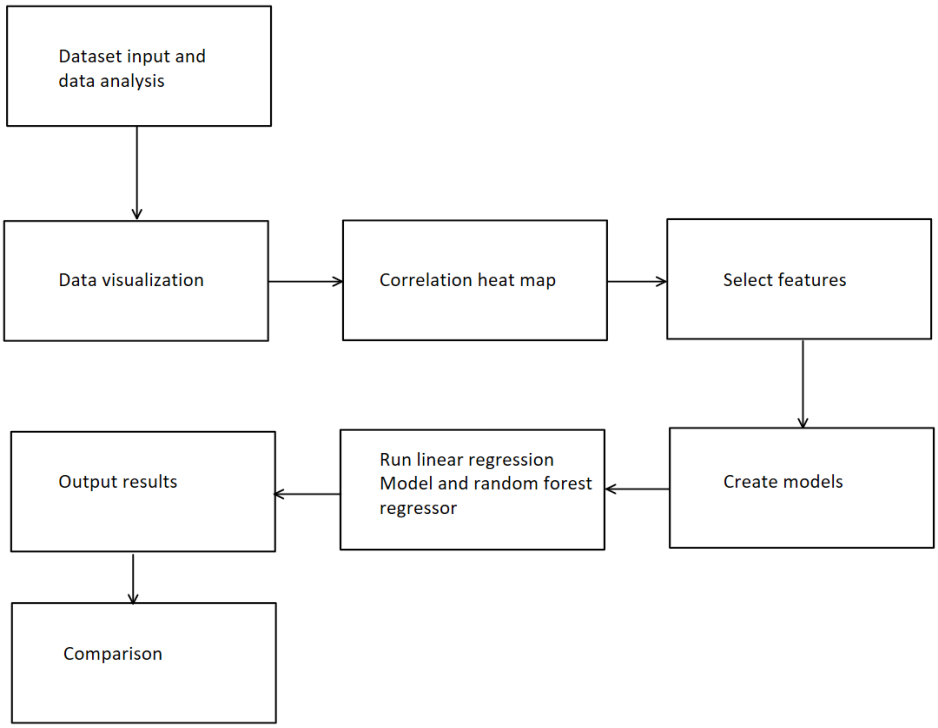

Figure 1. The pipeline of the study

2.2.1. Visualization analysis

Data visualization is used in order to explore and discover our data, to identify trends and patterns, and to better understand data types and trend changes before building models to inform future modeling. Visualization is done by drawing various icons such as histograms, line graphs, and scatter plots to get the relationship between the data and use them to filter the data. The data was first compared to position to differentiate the average data for different positions, as there are large differences in scoring, rebounding, etc., between different positions.

Then, we compare each data according to the age distribution of players and find that the influence of age stage on players' ability is not obvious. Finally, according to the summary of the above data, we can get the rough ability mapping of players in different positions, including rebounds, assists, steals and other aspects, to have an intuitive understanding and analysis of their value. Secondly, in order to avoid the problem of multiple covariance caused by multiple data in the model, we have analyzed the correlation of multiple data in the two datasets and drew a correlation heat map on this basis. Based on this map, we can quickly identify data with high correlation.

2.2.2. Logistic regression

Linear regression is a supervised machine learning algorithm used to calculate the linear relationship between a dependent variable and one or more independent features [11]. When the number of independent features exceeds one feature, it is called multiple linear regression and is used in our code snippet. Multiple linear regression is an extension of simple linear regression to the case of multiple independent variables. It is also a special case of the general linear model and is limited to one dependent variable [12]. The basic model of multiple linear regression is:

$$Y = \beta_0 + \beta_1 x_{i1} + \beta x_{i2} + \beta_3 x_{i3} + \cdots + \varepsilon \quad (1)$$

In the above formula, we consider n observations of one dependent variable and p independent variables. Therefore, Yi is the i-th observation of the dependent variable, and Xij is the i-th observation of the j-th independent variable, j = 1, 2, ..., p. βj represents the parameter to be estimated, and εi is the i-th independent and identically distributed normal error [12].

2.2.3. Random Forest

Random Forest Regression model requires hyperparameter tuning to optimize its performance. The goal is to find the best combination of hyperparameters to achieve the highest R2 score on the test data. A list of test sizes and random states is generated. In a nested loop, the dataset is repeatedly split into training and test sets using different test sizes and random states. For each combination of test size and random state, the Random Forest Regression model is instantiated and specific hyperparameters are set, including 100 estimators and a maximum depth of 5. The model is trained on the training data. Predictions are made on the test data, and the R2 score is calculated. Random Forest is a versatile ensemble learning technique that consists of multiple decision trees, each trained on random samples of the data and random subsets of features. It excels in predictive stability by integrating the results of these decision trees, thus improving model accuracy and robustness. By randomly selecting samples and features and aggregating the predictions of multiple decision trees, Random Forest mitigates the risk of overfitting and enhances model generalizability. Additionally, Random Forest can evaluate feature importance, aiding in understanding which features are more influential in prediction results, making it suitable for various tasks, including climate prediction. Its parallelizable nature allows for efficient model training, contributing to its widespread adoption in data science and machine learning applications.

3. Results and Discussion

In this paragraph, we put our predicted data together and compare them with the real data so that we can see the visual comparison clearly. Four different types of plot are used for the Figures 2-5.

A scatter plot is generated to compare the actual points (x-axis) with the predicted points (y-axis). Each point is color-coded based on the actual points. The plot is created using Plotly's scatter plot function. The result is shown in the Figure 2. This shows that the actual predicted development direction is almost consistent with the true value trend, and the model has excellent predictive ability.

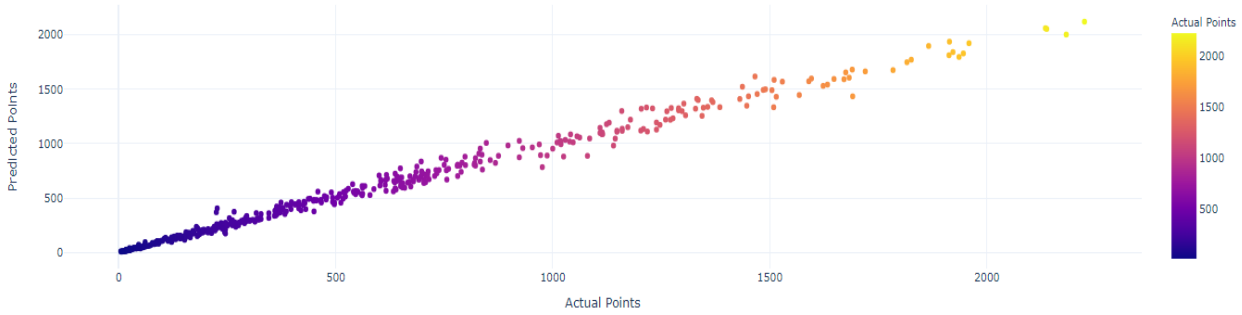

Figure 2. The comparison of actual and predicted result

A histogram plot is created to visualize the distribution of actual and predicted points, is shown in the Figure 3. Both distributions are overlaid on the same plot. The plot is generated using Plotly's histogram plot function.

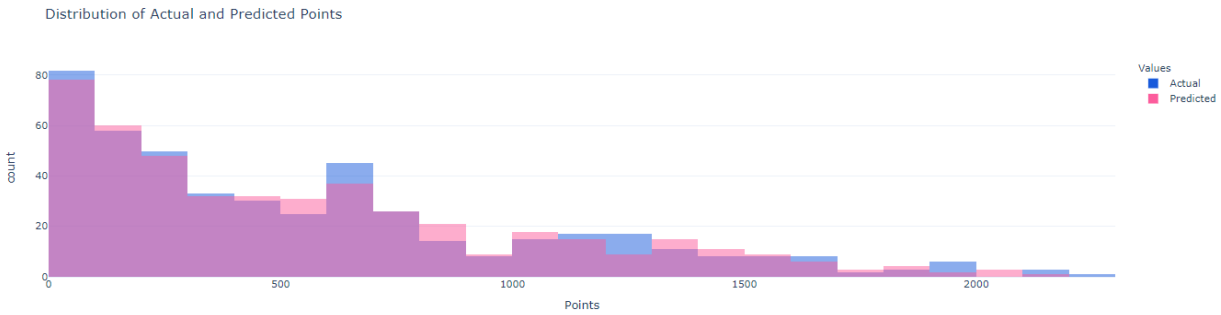

Figure 3. The distribution of actual and predicted points

A residual plot is generated to display the differences between the actual points and the predicted points, which is shown in the Figure 4. The residuals are calculated as the difference between the actual and predicted points. A dashed orange line at y=0 helps in visualizing the deviation from the ideal line. The plot is created using Plotly's scatter plot function.

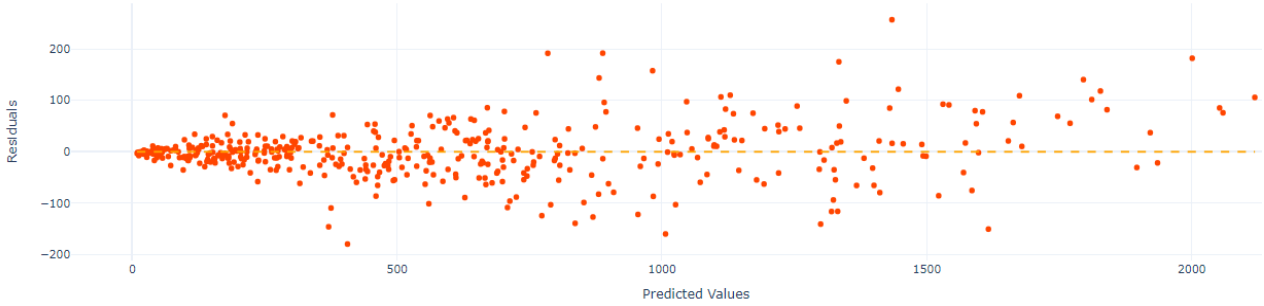

Figure 4. The residual plot

This plot depicts a comparison between the true values (x-axis) and the predicted values (y-axis). An ideal line, regression line, and scatter plot of the predicted values are shown in the Figure 5. The regression line represents the linear relationship between the true and predicted values. The plot is generated using Plotly's scatter plot function.

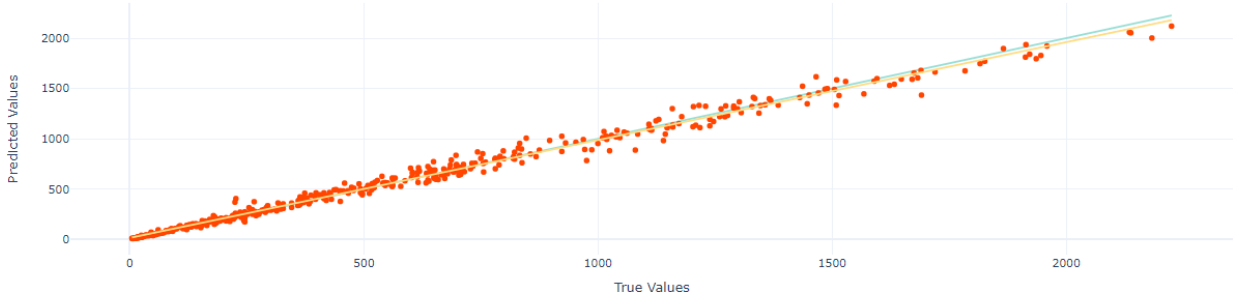

Figure 5. The predicted and the true line plot

4. Conclusion

In this paper, we present machine learning models aimed at constructing a comprehensive quality prediction model for forecasting NBA players' specific scores. The objective is to aid in the analysis, guidance, and evaluation of players' relevant value. Regarding data processing methods, we first rename the dataset and partition it into an 80% training set and a 20% dataset through preprocessing. Simultaneously, we address missing rows by assigning the position SG. We then visualize the overall data and conduct correlation analysis by group, generating a correlation heatmap to mitigate multicollinearity issues. Based on the visualization chart's summary, we delineate a tentative ability map of players across different positions, encompassing rebounds, assists, and other facets. Subsequently, employing random forest and linear regression methods, we harness NBA player data to train the model. We proceed to compare the performance of different models and analyze their respective strengths. Eventually, we derive histograms and linear graphs for the LR and RF models. Notably, RF demonstrates superior fitting to the data, whereas LR's predictions diverge from the ideal line, suggesting that the RF model yields more accurate predictions.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] V. Sarlis, C. Tjortjis, Sports analytics—Evaluation of basketball players and team performance, Information Systems, 93, 2020, p: 101562.

[2] K. Wheeler, Predicting NBA Player Performance, Wheeler-PredictingNBAPlayerPerformance, 2012.

[3] M. Casals, A. J. Martinez, Modelling player performance in basketball through mixed models, International Journal of performance analysis in sport, 13(1), 2013, pp: 64-82.

[4] C. Ge, et al, Predicting the outcome of NBA playoffs based on the maximum entropy principle, Entropy, 18(12), 2016, p: 450.

[5] L. Bernard, E. Bednar, and W. Bauer, Predicting NBA games using neural networks, Journal of Quantitative Analysis in Sports, 5(1), 2009.

[6] J. Lu, Y. Chen, and Y. Zhu, Prediction of future NBA games' point difference: A statistical modeling approach, 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), IEEE, 2019.

[7] F. Paul, and B. Taylor, On estimating the ability of NBA players, Journal of Quantitative analysis in sports, 7(3), 2011.

[8] Z. Albrecht, Basketball predictions in the NCAAB and NBA: Similarities and differences, Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(5), 2016, pp: 350-364.

[9] C. Bryan, et al, Predicting the Betting Line in NBA Games, 2013.

[10] Information on:https://www.kaggle.com/code/amirhosseinmirzaie/nba-players-scored-points-prediction/notebook.

[11] Information on: https://www.geeksforgeeks.org/ml-linear-regression/

[12] Information on: https://en.wikipedia.org/wiki/General_linear_model