NBA Player Score Prediction Based on Machine Learning

Haoyu Chen
Faculty of Science, China Pharmaceutical University, Nanjing, China

Keywords: Machine Learning, Visualization Technology, Random Forest, Linear Regression.

Abstract: With the success of machine learning and data visualization in many fields, the NBA(National Basketball Association) has also benefited from its huge demand for data analysis. These analysis results have been extensively applied in player draft, player training and tactical decisions, playing a crucial role in management and coaching staff decisions. This article utilizes data visualization technology and machine learning to analyze the NBA dataset. Using random forest and multiple linear regression models to predict NBA player scoring performance, and evaluate the model using R-square scores and MAE(Mean Absolute Error). There are some significant relationships between Points and several features like Turnovers, FGM and Minutes Played. After a ten-fold validation experiment, it was found that both the multiple linear regression and random forest are greater than 0.98 in R-square scores. And according to the result of the comparison, the multiple linear regression model is more suitable as a score prediction model and has a better stability for this dataset.

1 INTRODUCTION

Basketball is one of the most popular sports in the world. Many people are attracted by its entertainment and antagonism. Therefore, NBA (National Basketball Association), the top basketball league in the world was focused on millions of fans worldwide who eagerly followed the performances of their favorite teams and players. Behind the scenes, teams and coaching staffs have recognized the value of leveraging data to gain a competitive edge. By extracting insights from vast amounts of historical game data, teams can optimize strategies, enhance player performance, and make informed decisions both on and off the court (Thabtah et al 2019). Team managers in the NBA are beginning to gradually realize the huge potential of data analysis in basketball and are attempting to recruit data analysts to carry out further quantitative analyses of their players' physical condition and game performance. However most traditional data analysis only uses tools such as line charts to visually present players' various data or uses heat maps to display players' sweet zones (Georgievski and Vrtagic 2021). Such methods can only have a superficial understanding of the data set as it can not find the interaction among various features. Otherwise, the noise hiding behind the features can seriously influence the result of analysis. The traditional analysis methods can not solve these problems. So machine learning fills this area perfectly. With the development of machine learning, data analysts can utilize the rich and diverse dataset to offer more comprehensive and three-dimensional analysis, including regression, classification and clustering. This paper will predict player scores by using machine learning and statistical analysis based on a dataset from Kaggle which contains ample samples of NBA players. The dataset would be presented by several charts and tables visually. After analyzing the charts and pictures obtained from the result of data preprocessing and visualization, the p-value and student test would be used to determine the final features. These features should have greatly impression on the model. Then the methods of linear regression and random forest would be used to predict scores based on these selected features. Linear regression is a mathematical and statistical method that determines the parameters of a straight line by examining the relationship between the independent and dependent variables to find a line that best fits all samples. Predictions are made on new data through the model obtained. Random forest is an algorithm based on decision trees. It builds multiple different decision trees by picking features multiple times. Predictions on new data would gain a final prediction

Chen, H.
NBA Player Score Prediction Based on Machine Learning.
DOI: 10.5220/0012801700003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 291-296
ISBN: 978-989-758-705-4

based on the results of these decision trees. The fitting model should have ability to make prediction on NBA player scores based on new data. The final results can offer some help to managers and coaching staff to judge a player's potential and ability. Then they can make decisions more scientific.

2 EASE DATA COLLECTION AND PREPROCESSING

2.1 Data Source

The dataset used in this project is from Kaggle, the world's largest data science community. This dataset describes the data of all NBA players in the 2022-2023 season which contains 539 samples and 29 features.

2.2 Preprocessing

Table 1: Dataset information.

	Column	Non-null Count
1	Player	539
2	POS	534
3	Team	539
4	Age	539
...	...	...
29	Triple Doubles	539

Plenty of data samples have been used in this project, it would be appropriate to check the original data. It can be seen from Table 1 that this dataset has 29 features, Player, Position, Team, Age, Games played, Win, Lose, Minutes, Points, Field Goal Made, Field Goal Attempt, Field Goal Percentage, 3 points Made,3 points Attempt,3 points Percentage, Free Throw Made, Free Throw Attempt, Free Throw Percentage, Offensive Rebounds, Defensive Rebounds, Assists, Turnovers, Steals, Blocks, Personal Fouls, Fantasy Points, Double Doubles, Triple Doubles.

From Table 1, it can be seen that every feature has no missing values except Position. This feature only has 534 non-blank samples which means there are 5 missing values. From detailed information, it can be found that the longest playing time of these five players in the 2022-2023 season does not exceed 24 minutes as Table 2. This phenomenon means that these five players were signed by the team for the purpose of filling salary and personnel vacancies, and they mostly played for the development league teams under their team or only had a ten-day short contract. Their statistics are little useful to this project, so they would be dropped. And transform the column names to make them more readable, so that it can ultimately obtain the required dataset. This dataset has 534 samples.

3 DATA VISUALIZATION SECTION

After obtaining the dataset required, it would be appropriate if Data Visualization Analysis technology is applied to the statistics. Data Visualization Analysis technology can extract complex numerical values

Table 2: Detail information.

	Player	POS	Team	Age	GP	…	Min	PTS
1	Jason Tatum	SF	BOS	25.0	74.0	…	2732.2	2225.0
2	Joel Embiid	C	PHI	29.0	66.0	…	2284.1	2183.0
3	Luka Doncic	PG	DAL	24.0	66.0	…	2390.5	2138.0
4	Stephen Curry	PG	GSW	35.0	56.0	…	1941.2	1648.0
…	…	…	…	…	…	…	…	…
533	Alondes Williams	NAN	BKN	23.0	1.0	…	5.3	0.0
534	Deonte Burton	NAN	SAC	29.0	2.0	…	6.5	0.0
535	Frank Jackson	NAN	UTA	24.0	1.0	…	5.0	0.0
536	Michael Foster Jr.	NAN	PHI	20.0	1.0	…	1.0	0.0
537	Sterling Brown	NAN	LAL	28.0	4.0	…	24.4	0.0

from datasets and present them using figures or other methods, making them clear at a glance (Mingrui 2023). This technology has improved the flexibility, readability, and operability of datasets. And make it possible for data analysts to have a new method to operate the statistics and lay a solid foundation for further study or operation of the datasets (Wang and Fan 2021).

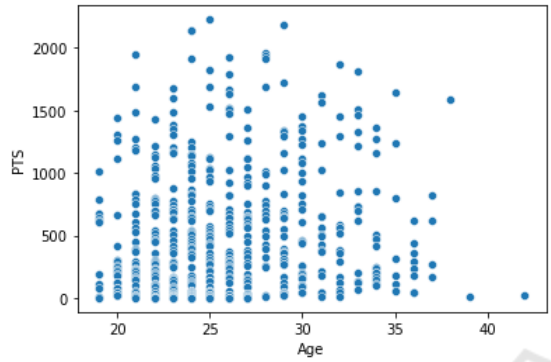

Figure 1: Relationship between PTS and age.

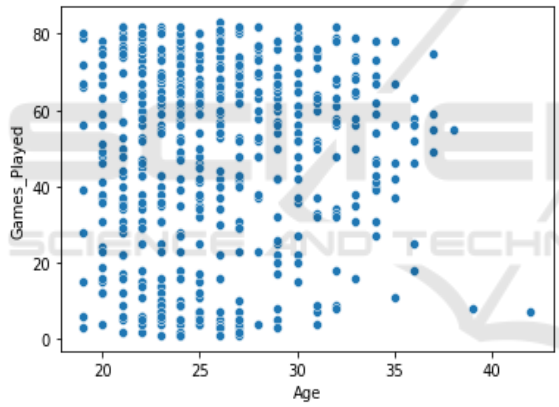

Figure 2: Relationship between games played and age (Picture credit: Original).

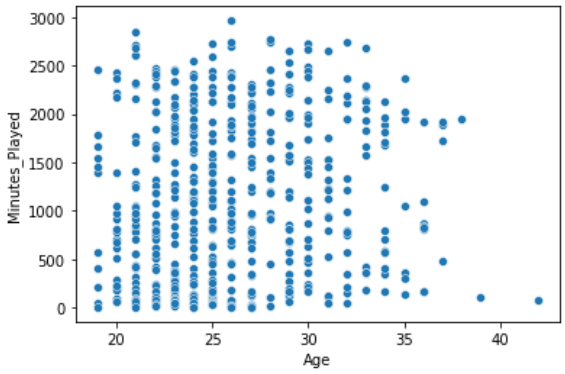

Figure 3: Relationship between minutes played and age (Picture credit: Original).

In Figure 1 it reveals that the players aged under 20 and beyond 35 have lower scores in total. The young players just enter the league and start their careers. Their body, mind and skills have not prepared enough to meet the new challenge. When it comes to the olders, Their numbers are in the minority of the entire league. According to some data, most of the NBA players would be retired at about 34 years old (Yongrong and Xiaojuan 2019). Their physical abilities will decline after the age of 33, unable to support them for a whole season of intense competition. It is common for older all-star players to choose to rest in the unimportant games to maintain their bodies for the long run. But for the common players, if their body can not afford to serve their team, it is difficult for them to obtain a contract. They have to turn to other leagues search for a job or just retire. This is also well reflected in Figure 1, there are only 3 players beyond 37 years old in 2022-2023 season. Older players have played fewer games and minutes, so their total scores are not as good as young players as shown in Figure 2 and Figure 3.

Table 3: Correlation between pts and other features.

	PTS
Age	0.092
GP	0.71
W	0.66
L	0.58
MIN	0.91
FGM	0.99
FGA	0.99
FGP	0.15
3PM	0.77
3PA	0.79
3P%	0.16
FTM	0.92
FTA	0.91
FT%	0.31
OREB	0.48
DREB	0.79
REB	0.74
ASSISTS	0.79
TOV	0.92
STL	0.77
BLK	0.47
PF	0.75
FP	0.97
DD	0.59
TD	0.27
+/-	0.28

And the majority of players with high scores are at the age range from 25 to 30. It seems that Age is a significant feature that can influence the score. The heat map can also be used to research the relationship

between features. It can be seen from Table 3 that the values of Minutes Played, Field Goal Made, Field Goal Attempted, Free Throw Made, Free Throw Attempted, and Turnovers in the PTS row are all more than 0.9. They have a high correlation with PTS.

4 PREDICTION ON NBA PLAYER SCORES

4.1 Method Introduction

In this study, the machine learning method would be used to make predictions on NBA players. This technology has demonstrated strong capabilities in data classification, prediction, and other aspects. To reach the goal, Random Forest and Multiple Linear Regression would be used.

4.1.1 Random Forest

There is a method in enterprise management called brainstorming, which can combine the wisdom of everyone to find solutions to problems. Not to come singly but in pairs, scientists applied this idea to machine learning, and ensemble learning was born. Ensemble learning refers to the learning of multiple estimators through training. When prediction is needed, the results of multiple estimators can be combined through the combiner, and the combined results can be output as the final result. In Random Forest, the estimators are the decision trees. The decision trees are combined together and the selection of the training dataset is totally random. Every time the part of the features would be chosen to enter the model. This process is totally random too. It is the reason that this algorithm is called Random Forest. When it tries to solve problems about regression, it would choose the mean of multiple results as the final result (Oughali et al 2019 & Young et al 2020).

4.1.2 Multiple Linear Regression

Multiple Linear Regression is a statistical analysis method. It is a kind of expand to Linear Regression to consider the impact of multiple independent variables on the dependent variable. A dependent variable is supposed to share some linear relationship with several independent variables. The model will calculate the values in datasets to find if there is a best-fit line exists. This line is of great help in the prediction of the dependent variables based on the given independent variables. The mathematical expression for multiple independent variables in multiple linear regression is

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\varepsilon \quad (1)$$

Y represents the value of a dependent variable. X represents several independent variables. β means regression coefficients. ε means the error term. The regression coefficients can explain the level of influence on the dependent variable by independent variables. This model will use the least squares estimator to find the most suitable regression coefficients. Because there is not much error between the predicted value of the model and the actual value, it is very common and useful to apply this model to a wide range of fields making predictions (Weihao and Jiliang 2022, Shi et al 2022, Kai and Jinhuan 2019 & Shun 2021).

4.2 Feature Selections

Table 4: Feature selection.

	t	p>|t|
CONST	-1.806	0.071
GP	0.401	0.688
MIN	6.390	0.000
FGM	27.141	0.000
TOV	-2.127	0.034
FGA	12.459	0.000
FGP	1.720	0.086
FTM	14.944	0.000
FTA	-5.209	0.000
OREB	-6.488	0.000
DREB	-0.194	0.846
ASSIST	-3.307	0.001
STL	-1.539	0.124
BLK	0.116	0.908
PF	-0.524	0.600
DD	1.488	0.137
TD	2.878	0.004

Setting all numerical features as independent variables and PTS as dependent variables. It is shown in Table 4 that the p-value of some features is very low which are under 0.05 while the value of t in these features is significantly high. The value of t is the ratio of the regression coefficient to its standard error. A larger absolute value of t indicates a more significant regression coefficient. And p value means the probability of the two-side hypothesis test of the t value. It represents that the observed difference between the regression coefficient and zero is caused by randomness. If the p-value is less than 0.05, it can be recognized that the regression coefficient is significantly not equal to zero by 95%. By using these

theories, Minutes Played, Field Goals Made, Turnover, Field Goals Attempted, Free Throw Made, Free Throw Attempted, Offensive Rebounds, assists, and triple-doubles are chosen to be the independent variables finally.

4.3 Result Analysis

This paper uses Python language to process the dataset. Dividing the dataset into two parts, the training set accounts for 80% while the 20% of samples in the test set would be used to evaluate the quality of the model. Training the random forest and linear regression models separately, and test them with the test set to obtain Table 5.

Table 5: R-square and MAE of models.

	R-square Score	MAE
Linear Regression	0.991	33.830
Random Forest	0.998	17.892

Table 6: R-square and MAE of models.

	Random Forest	Linear Regression
R-square Score	0.997	0.998
	0.993	0.998
	0.994	0.999
	0.997	0.997
	0.992	0.997
	0.991	0.994
	0.995	0.999
	0.985	0.999
	0.992	0.999
	0.992	0.999

The R-square score is a value used to evaluate the goodness of fit to a model. It represents the proportion of dependent variable variation that can be explained by the independent variables. The value range is between 0 to 1. The number is closer to 1, the better the model fits the observed data. Taking the average of the absolute value of the difference between each observed value and the predicted value as Mean Absolute Error (MAE). The smaller this value, the closer the prediction value to the observed value is. It is clear that the two models can perfectly make predictions on the PTS since the R-square scores are both beyond 0.99. But no matter in the R-square Score or MAE, the Linear Regression model performs better than the Random Forest. Furthermore, a 10-fold cross-validation method is used to evaluate two models as Table 6. The Linear Regression model is much more stable than the Random Forest model. Each score is closer to 1 and all are above 0.994. Although Random Forest performs well, the value is under 0.99 once. So in the prediction of NBA player scores using machine learning based on this dataset, Linear Regression is better.

5 CONCLUSION

In this study, the author uses data visualization technology and machine learning to analyze NBA datasets. By dividing the dataset into training and testing sets, the model makes predictions on NBA players' scores. The Multiple Linear Regression and Random Forest methods have been used. These two models are both performed perfectly in the prediction of NBA player scores. The R-square scores are above 0.99, which means models can describe and explain the data excellently. And Multiple Linear Regression performs better than Random Forest. There are some noises in the dataset that may cause this. Because many reasons can influence the performance of player scoring. For example, a player who just returned from injury can suffer from scoring. But this is not his true level and his scoring performance will return to normal levels as the season progresses. The random forest model is a decision tree-based model that may be more affected by these noises. This project proves that machine learning has huge potential in the field of NBA statistical analysis. Due to time constraints, this study is unable to predict the scores of different player positions separately. The model of different position players may also be different because of the duty they shoulder on the court. Hoping that in the future, this project can use machine learning to classify players and predict scores in different positions for further research.

REFERENCES

F. Thabtah, L. Zhang, N. Abdelhamid, “NBA Game Result Prediction Using Feature Analysis and Machine Learning,” Ann. Data. Sci. vol. 6, pp. 103–116, 2019.

B. Georgievski, S. Vrtagic, “Machine learning and the NBA Game,” Journal of Physical Education and Sport, vol. 21 no. 6, pp. 3339-3343, 2021.

Z. Mingrui, “Research on NBA Data Visualization and Integrated Learning Model Prediction Methods,” Guilin University of Electronic Technology, 2023.
J. Wang, Q. Fan. “Application of machine learning on nba data sets,” Journal of Physics: Conference Series. IOP Publishing, vol. 1802, no. 3, pp. 032036, 2021.
J. Yongrong, C. Xiaojuan, “Analysis of the causal relationship between NBA star data and team wins and losses: Taking LeBron James as an example,” Journal of Hefei Normal University, vol. 37, no. 3, pp. 29-31, 2019.
M. S. Oughali, M. Bahloul, S.A. El Rahman. “Analysis of NBA players and shot prediction using random forest and XGBoost models,”2019 international conference on computer and information sciences (ICCIS). IEEE, pp. 1-5, 2019.
C. Young C, A. Koo, S. Gandhi, et al. “Final Project: NBA Fantasy Score Prediction,” 2020.
J. Weihao, L. Jiliang, “A Study on the Factors Influencing the Value of Professional Basketball Clubs Based on Factor and Regression Analysis,” Inner Mongolia Statistics, vol. 1, pp. 39-43, 2022
Z. Shi, M. Li, M. Wang, et al. “NPIPVis: A visualization system involving NBA visual analysis and integrated learning model prediction,” Virtual Reality and Intelligent Hardware, vol. 4, no. 5, pp. 444-458, 2022.
Z. Kai, Z. Jinhuan. “Regression Analysis on the Scoring Ability of NBA East and West Teams: Taking the 2017-2018 Season as an Example,” Sports Science and Technology Literature Bulletin, vol. 27, no. 3, pp. 70-71. 2019.
L. Shun, “Analysis of NBA Player Salaries and Technical Data Based on Multiple Regression,” Contemporary Sports Technology, vol. 11, no. 4, pp. 229-232, 2021.