Dean&Francis

ISSN 2959-6157

The Key Indicators Affecting the Salaries of NBA Players are Analyzed Based on Stepwise Multiple Linear Regression and Random Forest Model

Xingwei Wang

College of Sciences, Shanghai University, Shanghai, China 20444
wangxingwei030910@shu.edu.cn

Abstract:

As the world's top-level basketball league, the National Basketball Association(NBA) has significant differences in player salaries, but the key influencing factors have not yet been fully clarified. Most of the existing studies focus on the linear relationship between salary and performance indicators, ignoring nonlinear effects or factors such as business value and rookie contracts. This study is based on NBA data from 2020 to 2025, eliminating star players and rookies to reduce bias. It adopts stepwise multiple linear regression (SMLR) and random Forest (RF) models to explore the determinants of salary. After SMLR solved the problem of variable collinearity, it showed that playing time, player influence assessment, and ball-handling offensive percentage (Usage Percentage(USG%)) were the main linear predictive indicators of salary. The model adjusted R^2 to 0.614, explaining 61.4% of the salary variation. The random forest model further reveals the influence of nonlinear factors such as age(AGE), which may be related to special contracts such as the Bird clause. Its test set R^2 reaches 0.664, and the prediction error is lower than that of SMLR, especially performing better in the medium and high salary range. Research shows that a player's actual contribution (MIN, PIE) and tactical status (USG%) are the core drivers of salary, and the random forest model has more advantages in capturing complex relationships. This research provides the team management with a basis for quantitatively evaluating the value of players and helps them optimize the team configuration.

Keywords: NBA players; stepwise multiple linear regression; random forest model.

1. Introduction

The NBA is the highest-level basketball league in the world and also the sports league with the highest average income in the world. There are over 500 basketball players in the NBA, who play a total of five positions on the court: point guard, shooting guard, small forward, power forward and center. The incomes of players in the NBA vary greatly. For instance, in 2024, Stephen Curry earned $51,915,615, while Kaiser Gates only received $35,389. Therefore, there are many possible potential factors that have affected the players‘ salaries. Statistical analysis provides a scientific method for studying players‘ salary issues. David pointed out the huge influence of statistical analysis in professional sports events. Data collection and analysis deeply affect the understanding of professional sports-related workers (such as team managers and coaches) in sports and help them improve the level of their teams [1]. Since 1997, the NBA has been conducting statistics on various data of players and uploading them to the official statistics website. Researchers have analyzed the salaries of NBA players and their performances on the court through statistical methods. Early studies based on regression models, such as Kevin J. iger in 2000, concluded that players‘ salaries were related to three basic data - points, rebounds, and assists. Points were correlated with rebounds [2]. Subsequent researchers improved this study in different aspects. Yang introduced more on-field data, such as WS (Victory Contribution value), FT (free throws), etc., to make the analysis results more accurate [3]. The Berkeley Sports Analysis team classified the players‘ salaries into four categories through the K-Nearest Neighbors Clustering(KNN) model and conducted separate studies on them [4]. Lu‘s research is currently the most comprehensive. It uses multiple models and rich on-court performance data to predict the income of players in different positions in the NBA, and concludes that the LASSO and elastic network models have the best prediction effects [5]. The above-mentioned research, through continuous improvement of statistical methods and models, provides deeper insights into the relationship between NBA players‘ salaries and performance.

This paper selects two models, namely multiple linear regression and random forest, for modeling. Through the data provided by the NBA official for evaluating various performances of players on the court, the key indicators affecting players‘ salaries are identified. It aims to help practitioners in related fields accurately assess the value of players and provide reasonable salaries.

2. Data Selection and Research Methods

2.1 Data Source and Description

The various data indicators of the players in this article are derived from the high-level data recorded on the NBA‘s official statistics website from 2020 to 2025 (averaged per game), which includes 17 variables related to salary, as shown in Table 1 [6, 7]. The salaries of the players are derived from all the players recorded on Hoopshype from 2020 to 2025, and the salary caps for 2020 to 2025 are from the BASKETBALL REFERENCE website [8].

Yang mentioned that commercial value would affect players‘ salaries [4]. The research by Scott Kaplan et al. shows that the absence of star players can have a huge impact on the ticket revenue of teams [9]. Lockie et al. pointed out that team managers tend to pay the premium brought by star players [10]. Therefore, this paper eliminates the star players in the NBA who obviously have an impact on the team‘s income to ensure that the influence of various data of the players on their salaries can be analyzed fairly. Meanwhile, Lockie et al. also pointed out that the salaries of rookies are often not linked to their performance on the court but directly related to their draft positions. Therefore, it is necessary to exclude rookie contracts [10].

Stanek‘s research introduced the salary cap of the NBA (divided into hard salary cap and soft salary cap, the former is the set threshold that the total salary of a team cannot exceed, and the latter is that a team can break through the salary cap limit through an exception clause, but it must comply with the regulations of the league), which is mainly used to ensure the balance of the league [11]. The salary cap soared from 26.9 million US dollars in the 1997-98 season to 140 million US dollars in the 2024-2025 season, which led to a rapid increase in the number of contracts signed by players each year. Therefore, it is more reasonable to convert players‘ salaries into the proportion of the salary cap. Meanwhile, Joao Vitor Rocha, through the cluster analysis of players‘ on-court data, concluded that there are three periods of different styles in the NBA, among which the same style has been from 2013 to the present [12].The indicators affecting salaries vary under different styles. Considering the introduction of more reasonable salary cap calculation rules in the league in 2017 and the impact of the pandemic on players‘ salaries during 2019-2021, this article only considers the data from 2021 to 2025.

Some star players in the NBA, as well as those under the age of 25, were excluded to ensure the impact of commercial value and rookie contracts on the results. Meanwhile, some players had too little playing time and games, and their on-court statistics were not representative. Players with an average playing time of less than 15 minutes per game were also excluded. The final data set obtained contains 1,097 players (the same player in different years is regarded as different players).

Table 1. Variable definition

Abbreviation	Variable name	Value range
AGE	Age	[25,40]
MIN	Playing time	[15.1,43.5]
OFFRTG	Offensive efficiency (points per 100 rounds)	[78.8,127.7]
DEFRTG	Defensive efficiency (points conceded per 100 possessions)	[88.7,132.3]
NETRTG	Net efficiency value (OFFRTG - DEFRTG)	[-38.8,29.8]
AST%	Assist rate (the percentage of goals scored by a teammate when a player is on the field and assisted by him/her)	[0,47.7]

Continue Table 1.

Abbreviation	Variable name	Value range
AST/TO	Assist-to-error ratio	[0,13.0]
ASTRATIO	Assist ratio (The proportion of assists to a player‘s possession of the ball)	[0,48.0]
OREB%	Offensive rebound rate (the percentage of offensive rebounds grabbed)	[0,18.7]
DREB%	Defensive rebound rate (the percentage of defensive rebounds grabbed)	[2.2,33.6]
REB%	Total rebounding rate	[1.9,25.7]
TORATIO	Error rate (the proportion of errors in possession of the ball)	[0,50.0]
EFG%	Effective shooting percentage (corrected three-point weighted shooting percentage)	[0,83.3]
TS%	True shooting percentage (considering the all-round scoring efficiency of shooting and free throws)	[0,83.3]
USG%	Ball-handling offensive rate	[4.0,38.7]
PACE	The pace of the game (the number of rounds per 48 minutes)	[94.8,107.2]
PIE	Player influence assessment (Comprehensive indicators for evaluating players‘ impact on the game)	[-3.7,23.0]
POSS	The number of offensive plays participated by the player	[33.0,6200.0]
S/C	Player salary *100/ annual salary cap	[0.012,40.7]

2.2 Research Method

2.2.1 Stepwise multiple linear regression(SMLR)

Multiple linear regression is used to study the relationship between a dependent variable y With multiple independent variables $\left(x_1, x_2, \ldots, x_n\right)$,Its expression is

$$y=\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n+\grave{o} \quad (1)$$

where $\beta_1, \beta_2, \ldots, \beta_n$ is the regression coefficient, Error term ò Satisfy the normal distribution. After obtaining the data set, the regression coefficients are determined by the least square method.

Since there are many variables in this case, when conducting the correlation test, it was found that there were collinear relationships among many variables. Therefore, stepwise multiple regression (SMLR) was adopted for analysis to eliminate the influence of collinearity on the regression results. SMLR is a modeling method that iteratively selects the optimal predictor variable. The core idea is to gradually screen the explanatory variables based on the statistical significance criterion by combining forward selection and backward elimination. This method starts from the zero model first. In each iteration, the most significant variables are included in the model based on the preset significance level (the forward step), and at the same time, it checks whether the included variables lose their significance due to the addition of new variables (the backward step). If the p-value of a variable exceeds the exclusion threshold, it is removed from the model. This cycle repeats until no variable meets the inclusion or exclusion criteria. This dynamic screening mechanism can effectively solve the problem of multicollinearity and automatically determine the optimal combination of explanatory variables. The final generated model not only ensures the prediction accuracy but also avoids overfitting.

2.2.2 Random forest

The Random Forest Algorithm (RF) is an ensemble learning model composed of multiple decision trees. The random forest model consists of three steps, as shown in Fig. 1. First, randomly extract data from the original training dataset to form different sub-datasets; Then, for each sub-dataset, different feature parameters are randomly selected to train a decision tree respectively; Finally, based on the prediction results of all decision trees and the type of problem to be solved, the final predicted value is obtained by calculating the mode or average. Compared with the decision tree model, since the RF model randomly extracts samples and features, it can effectively avoid the overfitting problem and thereby improve the generalization ability. Furthermore, random forests can automatically handle missing values and nonlinear relationships, and can be used to capture the complex relationships among variables. It can also visually display the contribution of each variable to the prediction result through feature importance assessment.

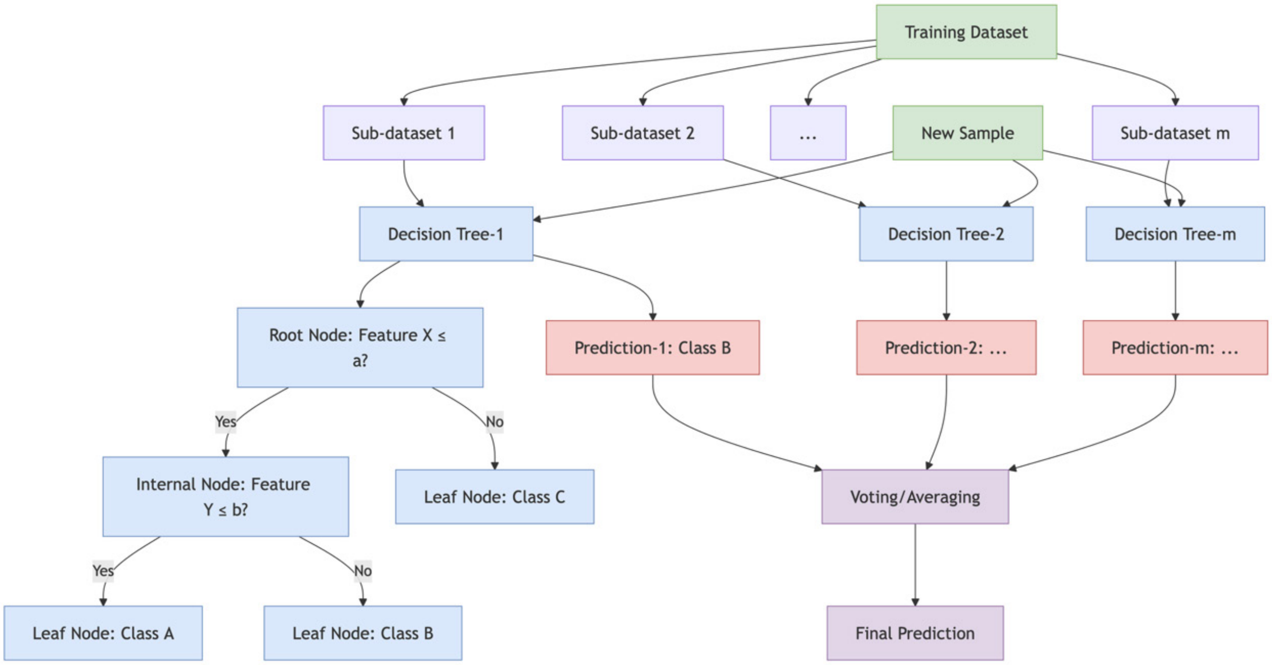

Fig. 1 Random forest flowchart (Original)

3. Results and Discussion

3.1 Stepwise Multiple Linear Regression

Since when conducting linear regression, it is necessary for the variables to have a significant linear relationship with S/C, and all variables in the dataset except AGE are continuous variables, it is possible to determine whether these variables have a linear relationship with S/C by calculating the Pearson correlation coefficient between each variable and testing their significance. The obtained result is shown in Fig. 2.

ISSN 2959-6157

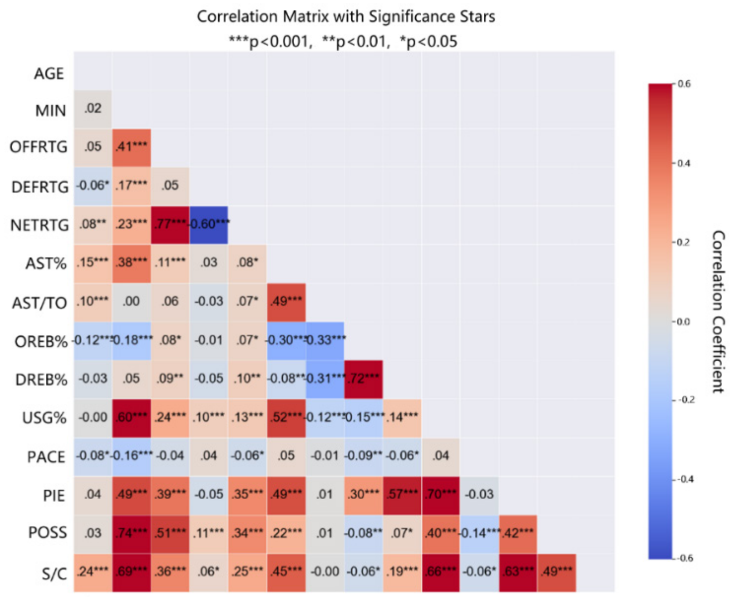

Fig. 2 Correlation coefficient heat map (Original)

It can be seen from Fig. 2 that MIN, OFFRTG, NETRTG, AST%, DREB%, USG%, PIE, POSS and S/C have significant linear relationships. Therefore, these variables are used as stepwise linear regression, and the results are shown in Table 2

Table 2. Results of Stepwise regression Analysis (n=1097)

	Regression coefficient	Significance coefficient p	VIF
Constant	-20.824	0.000**	\
MIN	0.733	0.000**	2.929
NETRTG	0.084	0.010*	1.265
AST%	0.074	0.005**	1.461
USG%	0.385	0.000**	2.604
PIE	0.688	0.000**	2.393
POSS	-0.000	0.020*	2.409
R^2	0.616		
Adjusted R^2	0.614		
Note: Dependent variable = S/C* p<0.05 ** p<0.01			

The calculation formula obtained thereby is

$$S/C = -20.824 + 0.733 * MIN + 0.084 * NETRTG \\ + 0.074 * AST\% + 0.385 * USG\% + 0.688 * PIE + 0.00 \\ * POSS$$

The VIF of all the parameters is less than 5, so there is no collinearity relationship and the conclusion is valid. The regression coefficients of MIN and PIE are 0.733 and 0.688 respectively, which are truly correlated with S/C and are the two largest among all regression coefficients. This indicates that long playing time and high player influence are significant characteristics of high-salary players. Meanwhile, the regression coefficient of USG% is also relatively high, at 0.385. USG% represents a player‘s ball-handling and offensive ability. This is in line with the current situation in the NBA that the ball-handling offense of each team is generally carried out by the top-ranked players in the team. The regression coefficients of NETRTG and AST% are 0.084 and 0.074 respectively. These two abilities can also positively increase players‘ salaries, but the increase is limited. This might be because the former is not only related to the players, but also to the overall ability of the team. The latter represents a player‘s assist ability. In today‘s leagues that emphasize offense, the contribution of AST%, which reflects assist ability, to salary is far less than that of USG%, which reflects offensive ability. However, POSS in Table 2 makes almost no contribution to salary, which indicates that the number of offensive plays a player participates in does not have a significant impact on salary. The adjusted R^2 is 0.614, indicating that this model can explain approximately 61% of the variables and the fitting effect is good.

3.2 Random Forest Model

Since not all data have a linear relationship with players‘ salaries, the random forest model can be used to better study the relationship between all variables and players‘ salaries. A total of four seasons from 2020 to 2024 were trained through the random forest model, and the data of 2025 was taken as the test set. Table 3 shows the model parameters adopted by the random forest in this study.

Table 3. Settings of random forest model

Parameter name	Parameter value
Data preprocessing	None
Training set ratio	0.8
The number of decision trees	100
Node splitting criterion	squared error
The minimum sample size of node splitting	2
The minimum sample size of leaf nodes	1
Maximum depth of the tree	不限制
Maximum feature number limit	auto
Has the sample been returned	Yes
Whether to conduct data tests outside the bag	Yes

The results obtained after operating the random forest model on the dataset are shown in Table 4

Table 4. Model evaluation results

Indicator	Training set	Test set
R-square value	0.957	0.664
Mean absolute error value MAE	1.568	4.147
Mean square Error MSE	4.487	31.202

Table 4 indicates that the random forest model can explain approximately 66% of the changes in S/C, and the average absolute error value between the predicted values and the true values obtained by this model is 4.147. Meanwhile, the random forest model can provide the contribution degree of each variable to the influence of S/C, that is, the importance, as shown in Fig. 3:

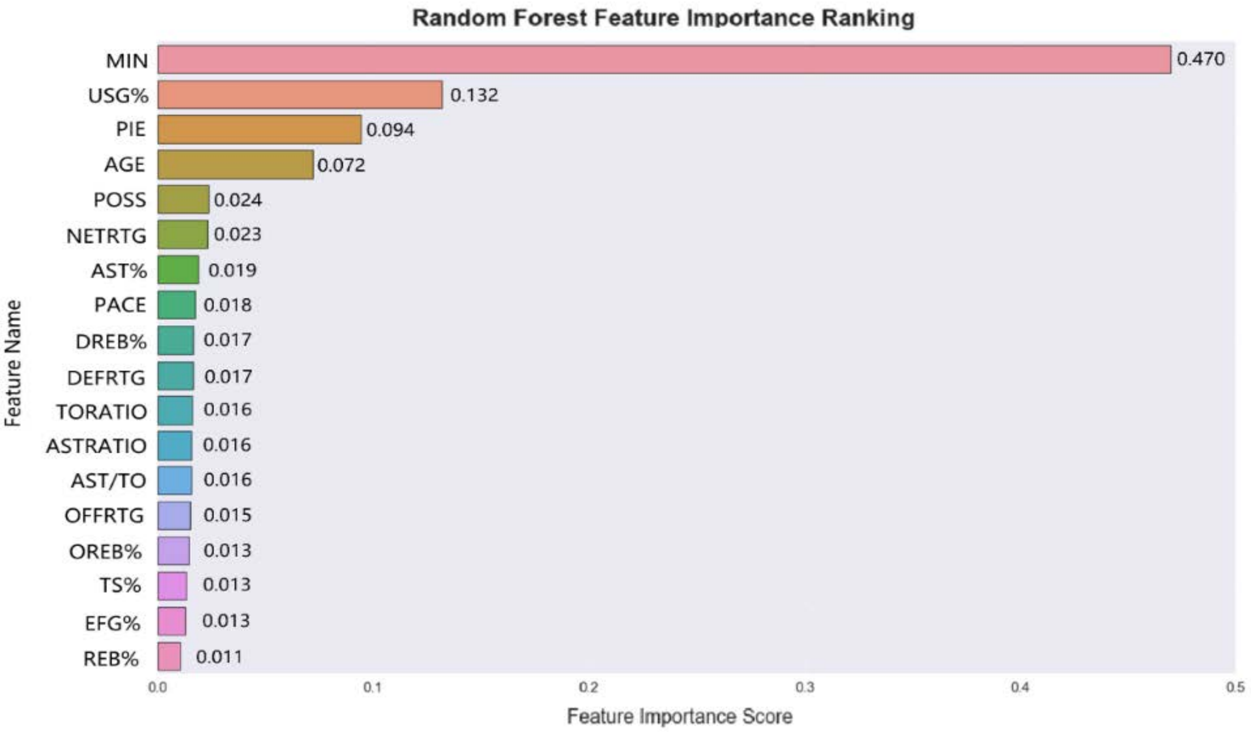

Fig. 3. Ranking of Importance of Various Indicators in Random Forests (Original)

MIN, USG%, and PIE all contribute significantly to the S/C contribution, just like in the stepwise linear regression model. The random forest model additionally points out that the AGE variable, which does not appear in the stepwise multiple linear regression, also has a considerable impact on S/ C. This might be because there are some terms in the league, such as the Larry Bird clause, that allow veterans to obtain higher contracts.

3.3 Model Comparison

Figs. 4 and 5, respectively present the salary predictions of NBA players in 2025 by stepwise multiple linear regression and the random forest model

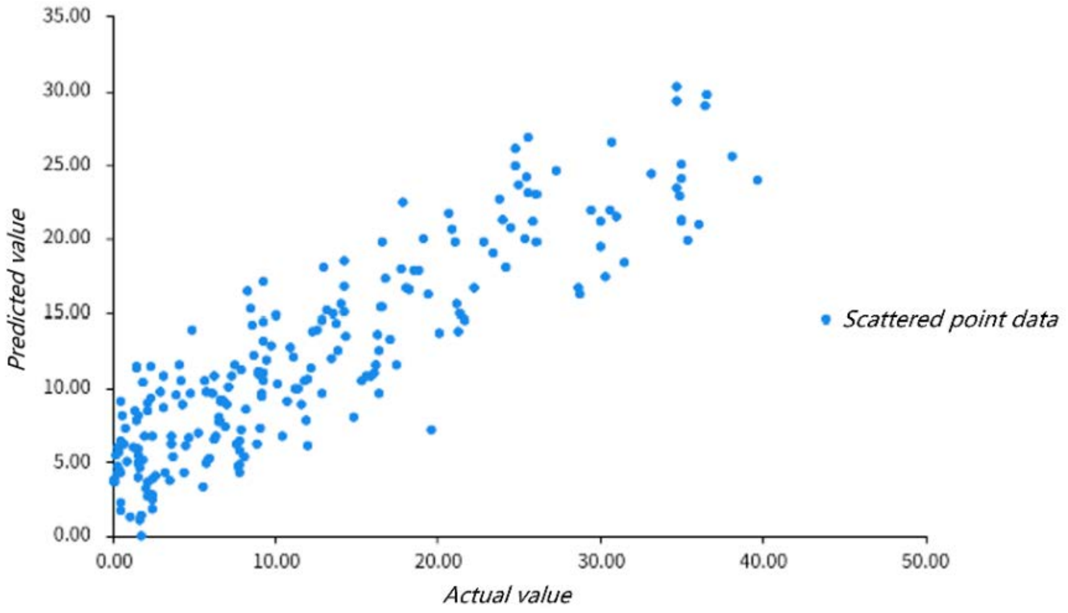

Fig. 4 Stepwise Multiple Linear Regression (Original)

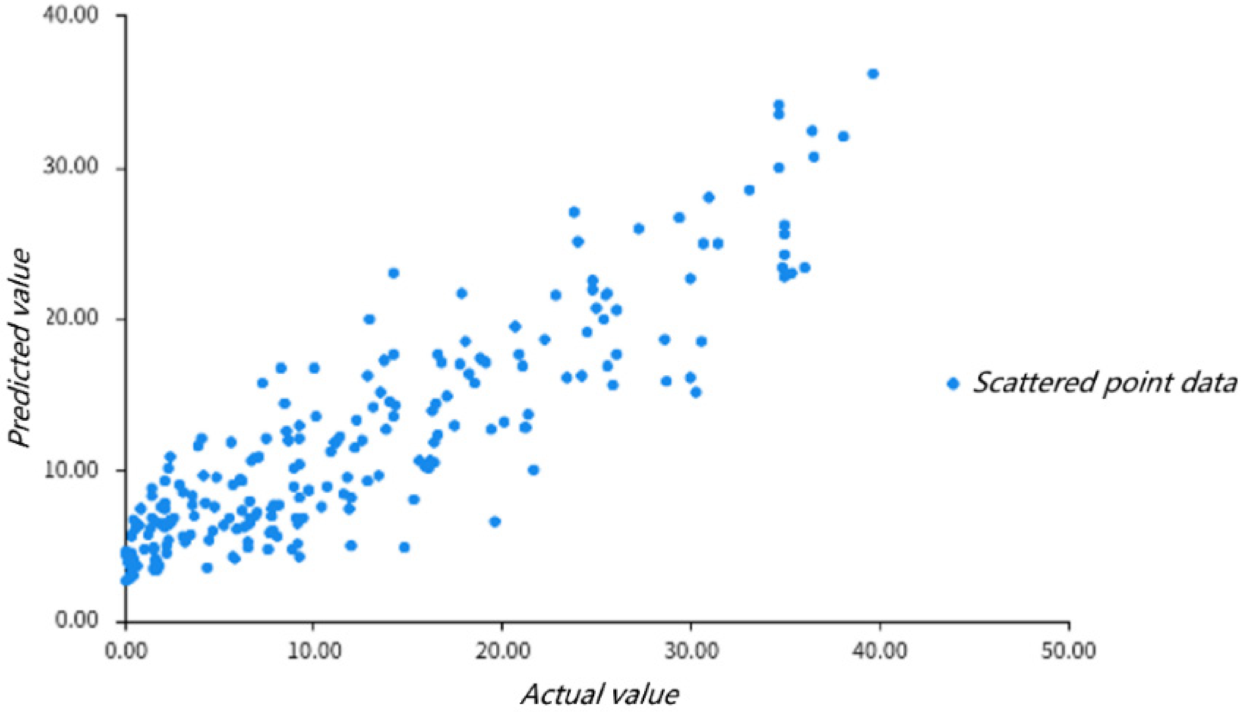

Fig. 5 Random forest (Original)

By comparing Fig. 4 and Fig. 5, it can be seen that the data point distribution of stepwise multiple linear regression is relatively scattered. Especially in the higher salary range (actual value > 30), there are prediction points that deviate significantly from the diagonal, indicating that the model has a large prediction error for high-salary players. Linear regression may have difficulty in capturing the nonlinear characteristics of salary. The data points of the random forest are distributed more closely around the diagonal. Especially in the medium salary range (actual value 10-30), the predictions are more concentrated, indicating that it can better fit the actual data. Therefore, in the problem of player salary prediction, random forest is a better prediction.

4. Conclusion

This study analyzed the key indicators affecting the salaries of NBA players through stepwise multiple linear regression and the random forest model. It was found that playing time (MIN), player influence assessment (PIE), and ball-handling offensive percentage (USG%) were the most important factors determining salaries. Among them, the regression coefficients of MIN and PIE were as high as 0.733 and 0.688 respectively. It indicates that the actual contribution and tactical position of players have a significant impact on salary. Furthermore, the random forest model further reveals the potential impact of AGE, which may be related to the special terms of the alliance. Model comparison shows that although stepwise regression can explain 61.4% of salary changes, it has a large prediction deviation in the high salary range. While the test set R^2 of random forest reaches 0.664, the prediction is more robust, especially suitable for players with medium salaries. These findings provide the team management with a basis for quantitatively assessing the value of players.

Future research can be further improved and deepened in multiple aspects: Firstly, further discussions on commercial value should be conducted, such as introducing data like the number of players‘ social media followers and commercial endorsement income. Meanwhile, the classification of players in different positions is discussed; Secondly, optimize the model methods, such as constructing a hybrid model combining linear and nonlinear elements, or introducing neural network models for analysis.

References

1. Yarrow D, Kranke M. The performativity of sports statistics: towards a research agenda. Journal of Cultural Economy, 2016, 9(5): 445-457.

2. Ioanna Papadaki and Michail Tsagris. Are NBA Players' salaries in accordance with their performance on court? Advances in Econometrics, Operational Research, Data Science and Actuarial Studies. Contributions to Economics, 2020.

3. Yang Z. Analysis of the Relationship between NBA Player Salary and Their On-Court Performances. Theoret-

ISSN 2959-6157

ical and Natural Science, 2024, 41: 51-58.
4. Wu W, Feng K, Li R, et al. Classification of NBA salaries through player statistics. Sports Analytics Group at Berkeley, 2018.
5. Lu Y. The Prediction of NBA Players' Salary. Advances in Economics, Management and Political Sciences, 2024, 57: 196-203.
6. NBA. 2025. https://www.nba.com/stats/players/traditional?SeasonType=Regular+Season&Season
7. NBA. 2025. https://www.nba.com/stats/players/advanced?SeasonType=Regular+Season&Season
8. Hoopshype. 2024/25 NBA Salaries. 2025. https://hoopshype.com/salaries/players/
9. Kaplan S, Ramamoorthy V, Gupte C, et al. The economic impact of NBA superstars: evidence from missed games using ticket microdata from a secondary marketplace. 13th Annual MIT Sloan Sports Analytics Conference. 2019: 1-30.
10. Lockie R G, Beljic A, Ducheny S C, et al. Relationships between playing time and selected NBA Combine test performance in Division I mid-major basketball players. International Journal of Exercise Science, 2020, 13(4): 583.
11. Stanek T. Player performance and team revenues: NBA player salary analysis. 2016.M. T. Lei, J. Monjardino, L. Mendes, D. Gonçalves and F. Ferreira, Air Quality, Atmosphere & Health, 2019, 12, 1049-1057.
12. Rocha da Silva J V, Rodrigues P C. The three Eras of the NBA regular seasons: Historical trend and success factors. Journal of Sports Analytics, 2021, 7(4): 263-275.