NBA Player Salary Analysis based on Multivariate Regression Analysis

Xuanhao Feng[1, *, †], Yinjie Wang[2, †] and Tianqin Xiong[3, †]

[1] School of Mathematical Sciences, Nankai University, Tianjin 300071, China

[2] School of Ocean and Earth Science, Tongji University, Shanghai 200092, China

[3] Sichuan University - Pittsburgh Institute, Sichuan University, Chengdu 610207, China

* Corresponding Author Email: 2012084@mail.nankai.edu.cn

†These authors contributed equally.

Abstract. The National Basketball Association (NBA) is one of the four major professional sports leagues in the United States and holds substantial economic value. Since the creation of the NBA, the issue of NBA player salaries has become increasingly relevant and the disproportionate salaries of players has emerged as a challenge that obstructs the growth of the league. This paper will employ data from 331 NBA players over the 2020-2021 season to develop multiple linear regression models for analyzing appropriate player salaries. The multiple regression analysis was conducted using basic and advanced data of players as independent variables, which include dozens of variables such as player's age, player's number of games, and player's average score, and player salary as dependent variable. In order to analyze the results more accurately, several types of multiple linear regression models are used in this paper: Ordinary least squares, Ridge regression, Lasso regression, Elastic net regression. The outcome of the regression model for NBA player salary will contribute to enhancing the commercial worth of the NBA and provide constructive input to the league and the team. This research found investors place higher significance on players' three-point shooting proficiency, defensive abilities, and teamwork capacity, which aligns with the current playing style of the NBA.

Keywords: NBA; player salary; OLS; lasso regression; elastic net regression.

1. Introduction

The National Basketball Association (NBA) has established itself as a highly lucrative and commercially valuable sporting event, featuring some of the world's top basketball players. The league is characterized by intense competition and high commercial appeal, and in recent years, investors have increasingly recognized the value of the NBA as a platform for investment. To capture a larger share of the sports market, the league aims to recruit the best players. Teams within the league must also recruit the best players at reasonable salaries to ensure winning percentages and ultimately maximize revenue for both the team and the league. Therefore, determining reasonable player salaries has become a crucial issue. Adopting a method for analyzing appropriate player salaries would serve as a valuable reference for improving the overall performance of the basketball league.

Player value is a critical aspect in the realm of sports, and the performance and operations of sports leagues are greatly influenced by player transactions. Players are considered the primary production factor within the professional sports industry, with player salaries being the most significant determinant of this [1]. The four major sports leagues in the United States (NBA, NFL, MLB, and NHL) all have factors that are considered when determining player salaries. Players can increase their value by developing specific skills that are relevant to their positions [2]. Athletes who wish to enhance their value will continue to cultivate and refine the corresponding sports skills. Teams that aim to acquire the most talented players will continually optimize their methods for evaluating player salaries. This creates a positive feedback loop that ultimately leads to the enhancement of the league.

The field of NBA player salary analysis is relatively under-researched, with studies primarily focusing on three main areas: market value and market analysis from a marketing perspective, examination of the league's checks and balances system from a sports economics perspective, and analysis of team performance, player offense and defense, and player salary from a technical perspective related to the game of basketball itself. However, there is a lack of research utilizing statistical methods in this area. This paper aims to address the third point from a statistical perspective, which is the analysis of team performance, player offense and defense, and player salary from a technical perspective. Chen Shen utilized multiple linear regression to analyze the correlation between the overall ability of NBA players, their age, and their development prospects, with the goal of determining their salary. The results of the study indicate a strong positive relationship between a player's overall ability and their salary, as well as a negative relationship between a player's developmental potential and their salary [3]. There are a variety of factors that contribute to determining a player's salary in the NBA, including but not limited to statistics such as points per game, rebounds per game, assists per game, and blocks per game [4, 5]. While there is a positive correlation between these statistics and a player's salary, it should be noted that the number of appearances and steals per game have a negative correlation with salary [6]. Therefore, the key focus of this paper is the variables that have an impact on player salary. This paper presents an extensive examination of potential variable factors to establish a regression model for analyzing player salary.

There is a significant body of research on player salary in various sports leagues such as soccer, baseball, rugby, and basketball. NBA presents a unique case as it has a higher average of points per game and a greater number of factors used to evaluate individual player ability [7]. This makes it more conducive to regression models. Among the mathematical models employed to analyze player salary in the NBA, the multiple regression model is the most widely used. This model takes player salary as the dependent variable and includes independent variables such as average points scored, number of rebounds, number of assists, and other statistics from a season. This model utilizes statistical analysis methods such as multiple regression analysis to study the relationship between player salaries, analyze the salary distribution of most players, and ultimately, aid in the rationalization of the overall salary structure of the league [8, 9]. Multiple linear regression models have limitations in their ability to accurately analyze the salary of individual players. Ye Liu conducted an empirical study to examine the relationship between on-field performance and player salary income using a tree regression model. The study used data from 359 players, including salary income, on-field performance, and personal characteristics. The analysis revealed that a weighted combination model, which combines two tree regression models, namely random forest regression and XGBoost regression, was more effective in predicting player salary income based on on-field performance [10]. Therefore, this paper first applies a classical OLS regression, utilizing factors known to influence player compensation as independent variables. In order to improve the accuracy of our analysis model, this paper adopts a novel approach by utilizing a Lasso regression model for player salaries, given the substantial number of independent variables [11].

This paper's objective is to examine the relationship between various factors and the salaries of NBA players in our data set, and subsequently, to find a more precise analysis model. To achieve this goal, based on the limitations of the traditional data, this paper incorporates higher order data for statistical analysis to enhance the accuracy of our models. Furthermore, this paper employs OLS regression to develop analysis models for player salaries. In order to better solve the collinearity problem, ridge regression, Lasso regression and Elastic net regression are also used to analyze the data. Finally, to assess the validity and reliability of our models, this paper conducts an additional validation set to evaluate the analysis power of our models. It can provide valuable insights for basketball clubs to make informed decisions.

2. Methods

2.1. Data Source

In order to ensure the authority and accuracy of the data source, this paper uses the sub-website of American Sports Reference Company - Basketball-Reference as the data source. The company has subsites for various sports events such as basketball, baseball, American football, etc. These websites try to take comprehensive statistics on sports data. In order to ensure the comprehensiveness of the data, the data acquisition includes all the stadium data of 500 players in the two seasons of 2020-2021 and 2021-2022 on the Basketball Reference. We eliminated some players with missing data, and finally obtained all the field data of 331 players.

2.2. Variable Description

Table 1 and 2 show the full name, symbol and description of the 45 variables used in the study.

Table 1. Name, Symbol and Description of Variables (traditional data)

Full Name	Symbol	Value Range
Position	Pos	N/A
Age	AGE	[19,40]
Team	Tm	N/A
Games	G	[1,72]
Games Started	GS	[0,72]
Minutes Played Per Game	MP	[3,37.6]
Field Goals Per Game	FG	[0.3,11.2]
Field Goal Attempts Per Game	FGA	[0.6,23]
Field Goal Percentage	FG%	[0.231,1]
3-Point Field Goals Per Game	3P	[0,5.3]
3-Point Field Goal Attempts Per Game	3PA	[0,12.7]
3-Point Field Goal Percentage	3P%	[0,1]
2-Point Field Goals Per Game	2P	[0.1,9.2]
2-Point Field Goal Attempts Per Game	2PA	[0.2,16.8]
2-Point Field Goal Percentage	2P%	[0.194,1]
Effective Field Goal Percentage	eFG%	[0.308,1]
Free Throws Per Game	FT	[0,9.2]
Free Throw Attempts Per Game	FTA	[0,10.7]
Free Throw Percentage	FT%	[0.444,1]
Offensive Rebounds Per Game	ORB	[0,4.7]
Defensive Rebounds Per Game	DRB	[0.4,10.1]
Total Rebounds Per Game	TRB	[0.5,14.3]
Assists Per Game	AST	[0,10.8]
Steals Per Game	STL	[0,2.1]
Blocks Per Game	BLK	[0,3.4]
Turnovers Per Game	TOV	[0,4]
Personal Fouls Per Game	PF	[0,3.8]
Points Per Game	PTS	[1.1,32]

Among them, the variables listed in Table 1 are the players' basic play data (obtained from specific records), while the variables listed in Table 2 are higher-order variables (based on the basic data, the efficiency, hit rate and other dimensions are all considered, and the real contribution is calculated). This study will explore the contribution of all the above variables to salary.

Table 2. Name, Symbol and Description of Variables (higher order data)

Full Name	Symbol	Value Range
Player Efficiency Rating	PER	[-0.9,54.6]
True Shooting Percentage	TS%	[0.308,1]
3-Point Attempt Rate	3PAr	[0,0.889]
Free Throw Attempt Rate	FTr	[0,1.111]
Offensive Rebound Percentage	ORB%	[0,19]
Defensive Rebound Percentage	DRB%	[3.9,37.5]

Total Rebound Percentage	TRB%	[2.9,26.7]
Assist Percentage	AST%	[0,44.4]
Steal Percentage	STL%	[0,7]
Block Percentage	BLK%	[0,8.8]
Turnover Percentage	TOV%	[0,31]
Usage Percentage	USG%	[7.2,35.3]
Offensive Win Shares	OWS	[-2.9,9.6]
Defensive Win Shares	DWS	[0,5.2]
Win Shares	WS	[-1.9,11.3]
Win Shares Per 48 Minutes	WS/48	[-0.099,0.475]
Offensive Box Plus/Minus	OBPM	[-7.8,24.1]
Defensive Box Plus/Minus	DBPM	[-2.7,8.2]
Box Plus/Minus	BPM	[-9.6,31.1]
Value over Replacement Player	VORP	[-1.7,5.8]

2.3. Mathematical Statistics Method

2.3.1 Multiple linear regression model

The relationship between the response variables y_i and the predictor variables x_i can be modeled through a linear regression, where there are p predictor variables $x_1, x_2, \cdots, x_p$ and a single response variable y_i.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

Among them, ε_i is random error item, $\beta_0, \beta_1, \cdots, \beta_p$ are regression coefficient, and this is called a multiple linear regression model.

2.3.2 Ordinary least squares (OLS)

The residual sum of squares (RSS) is introduced on the basis of the multiple linear regression model. The smaller the RSS, the shorter the straight-line distance between the predicted value and the actual value, and the better the fitting effect at the same time.

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - x_i^T b)^2 \tag{2}$$

Among them, y_i is actual value, and the b which minimizes the RSS through the least square method is the corresponding OLS coefficient estimate $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$.

2.3.3 Ridge regression

For the general linear regression model, in order to avoid large parameter values, the coefficient β can be estimated by increasing the L_2 penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$, so that $\hat{\beta}^{\mathrm{ridge}}$ be the smallest.

$$\hat{\beta}^{\mathrm{ridge}} = \mathrm{argmin}_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

Among them, the larger the value of λ, the larger the shrinkage of the regression coefficient, the more accurate the coefficient estimate, the smaller the RMSE value. In particular, when $\lambda = 0$, ridge regression degenerates into OLS regression. However, because the ridge regression retains the coefficient estimates of all variables, this results in an unclear interpretation of the model.

2.3.4 Lasso regression

Before introducing the Lasso regression, the predictor variables as well as the response variables are standardized such that

$$\sum_{i=1}^{n} y_i = 0\,,\quad \sum_{i=1}^{n} x_{ij} = 0\,,\quad \sum_{i=1}^{n} x_{ij}^2 = 1\,,\quad j = 1,2,\cdots,n \tag{4}$$

Lasso regression is based on OLS with an additional L_1 penalty term to limit the length of the parameters. the estimator defined by Lasso is

$$\hat{\beta}^{\text{Lasso}} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (5)$$

In contrast to the penalty term $\lambda\sum_{j=1}^{p}\beta_j^2$ in ridge regression, the Lasso's L_1 penalty term $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ not only reduces the magnitude of coefficients β_j towards zero for non-zero predictor variables, but also identifies those predictor variables that have significant value (Predictor variables with large $\left|\beta_j\right|$ value).

2.3.5 Elastic net regression

Elastic net regression, like Lasso regression, necessitates the standardization of the data prior to analysis. In contrast to Ridge and Lasso regression, which only employ one penalty term each, Elastic net regression combines both L_1 and L_2 penalties to achieve improved fitting results.

$$\hat{\beta} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\beta_j^2 + \lambda_2\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (6)$$

If let $\lambda = \lambda_1 + \lambda_2, \alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, then

$$\hat{\beta} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}(\alpha\left|\beta_j\right| + (1-\alpha)\beta_j^2)\right\}, 0 \leq \alpha \leq 1 \quad (7)$$

3. Results and Discussion

3.1. Descriptive Analysis

As shown in Figure 1, it can be found that the span of player's salary is large, and its distribution is extremely right skewed. The median salary of players is only $5 million, while the highest salary of Stephen Curry is $45.78 million. The salary of most players in the NBA is less than 6.58 million dollars, and most of the players are far from the top of the pyramid with high income.

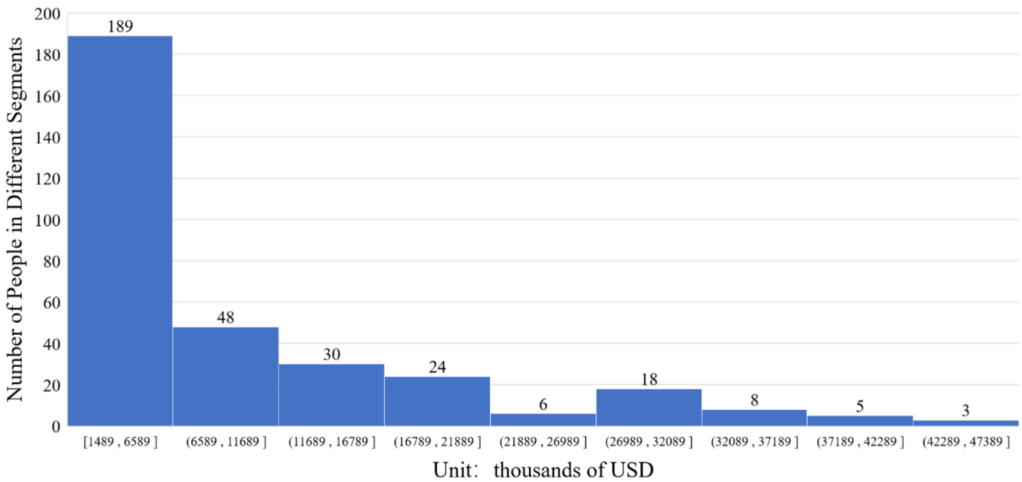

Fig. 1 Histogram of Player Value Distribution

Age is also an important factor that affects the salary of players. As shown in Figure 2, from the perspective of the salary of players of all ages, 25-32 years old is the golden age of players, and the average salary of players at this stage has reached the peak. The age of 30 will be regarded as a watershed. Before that, it is the growth period. The player's salary will rise with the age. After that, it began to enter a decline period, and the salary of players declined with the growth of age.

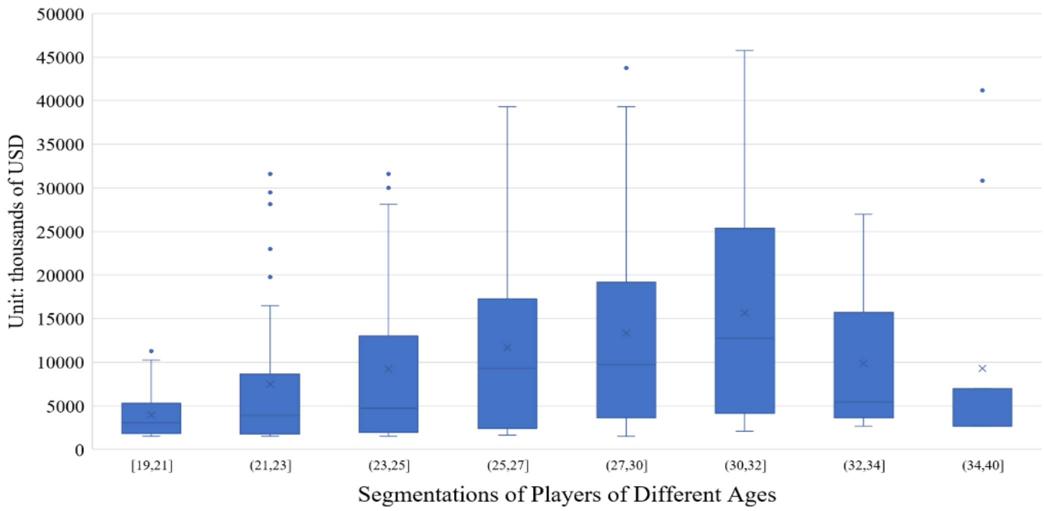

Fig. 2 Box Graph of Player's Salary at Different Ages

As for the relationship between positions on the field, there is no significant difference in the median salary of players in each position. However, in general, the salary of point guard is relatively the highest, the difference between point guard and power forward is not big, and the salary of point guard is the lowest. Some players have two positions in different situation, so this figure calculates their salaries twice on their different positions, as shown in figure 3.

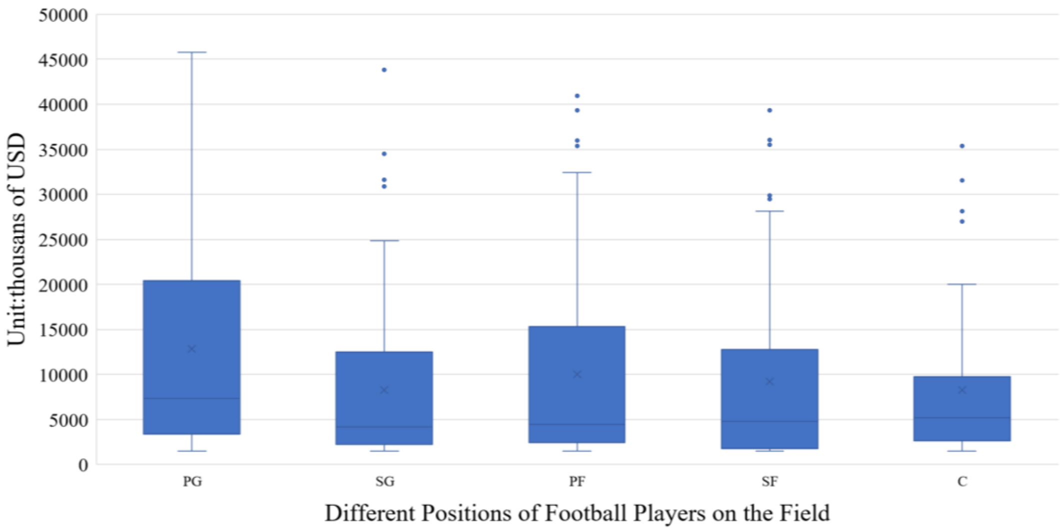

Fig. 3 Statistical Chart of Player Value Comparison in Different Positions

The game data often reflects the level of players' performance in the game. This paper obtains 45 game related data and carries out correlation coefficient analysis with the salary level from 2021-2022. According to the Figure 4, it can be found that many variables have strong correlation analysis. In particular, the average score per game (0.781), the average number of hits per game (0.759), the average number of shots per game (0.734) and the average number of free throws per game (0.734).

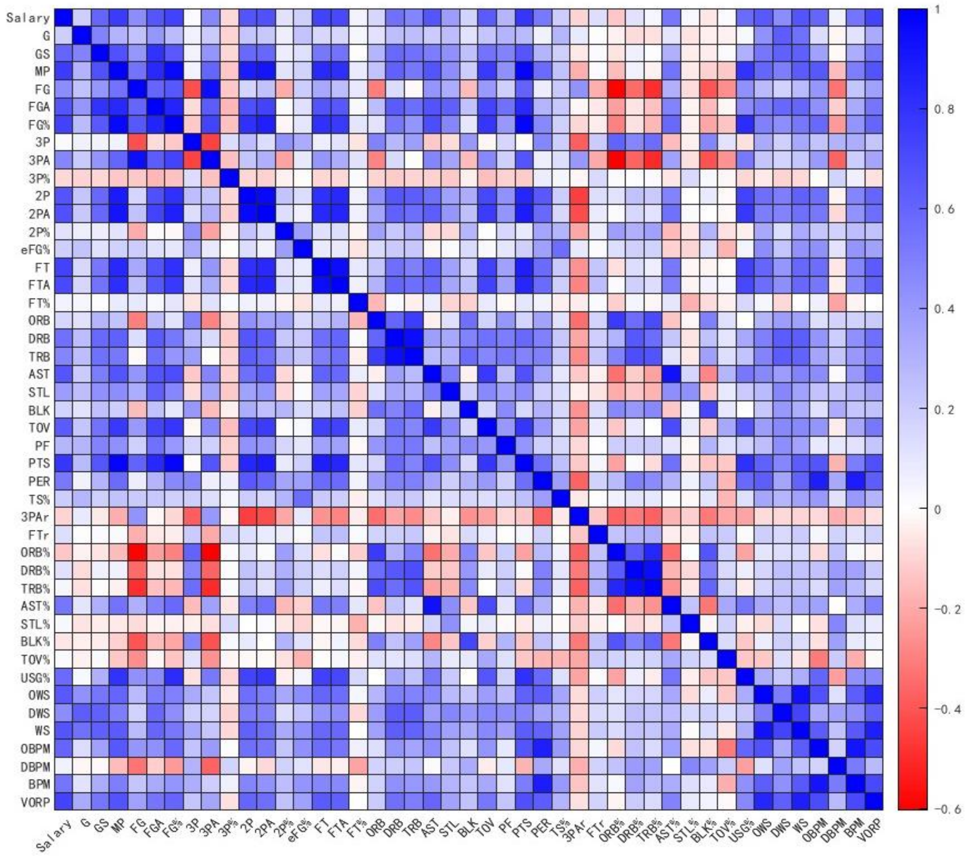

Fig. 4 Correlation Coefficient Diagram of 45 Factors Including Player Salary and Game Level

From Figure 4, we can infer that the higher the score, the higher the number of hits, the more shots, and the more free throw hits. Figure 5 shows the relationship between the four variables with the strongest correlation with salary.

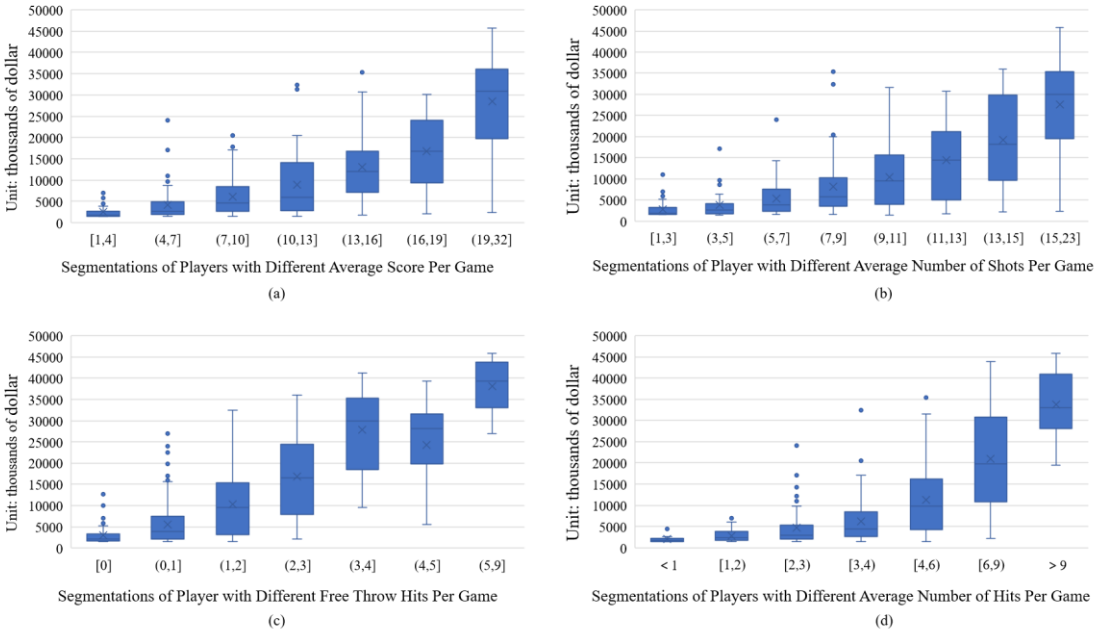

Fig. 5 Box graphs of the distribution of four representative variables with salary. (a) Salary Box Chart of Average Score of Players in Different Games (b) Salary Box Chart of Average Goals of Players in Different Games (c) Salary Box Chart of Average Number of Shots in Different Games (d) Salary Box Chart of Average Free Throw Hits in Different Games

3.2. Inferential Analysis

This paper believes that a player's salary is closely related to his performance in the previous year. Therefore, this paper selects all the players on the NBA player salary list in 2021-2022 and analyzes them with their data on the court in 2020-2021. In this paper, 257 people were taken as training samples, and the process was implemented by Python through random sampling. Then, the remaining 65 data samples were used to evaluate the quality of the model. The evaluation index was the root mean square error (RMSE) value, the smaller the value, the better the overall performance of the model. In the descriptive statistics, this paper found that the influence of age factor on players' salary is non-linear. With the increase of age, it increases first and then decreases. Therefore, the term $(\text{age} - 30)^2$ was changed for regression, and all data were standardized for subsequent regression.

Note that RMSE is a statistical measure used to evaluate the deviation between the predicted and true values of a variable. Specifically, it is calculated by taking the square root of the ratio between the squared difference of the predicted and true values, and the number of observations n. RMSE is recognized for its high sensitivity to outliers in the data, and it is often utilized in various research disciplines as a benchmarking tool for predictive models.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2} \tag{8}$$

First, the OLS coefficient estimation was obtained through programming in Python language, as shown in the second column of the Table 3 and Table 4. Next, Ridge, Lasso and Elastic Net methods were used to estimate, respectively. The results are shown in the Table 3 and Table 4.

Table 3. Regression coefficient (part 1: traditional data)

Variables	OLS	Ridge	Lasso	Elastic Net
AGE	-0.108	-0.094	-0.086	-0.11
G	-0.083	-0.07	-0.043	-0.079
GS	0.177	0.114	0.1	0.152
MP	-0.588	-0.024	0	-0.281
FG	-1.613	0.063	0	0.268
FGA	0.577	0.046	0	0
FG%	-0.031	-0.024	0	0
3P	-0.903	0.045	0	0.025
3PA	-0.431	0.049	0.005	0.064
3P%	-0.036	-0.045	0	-0.035
2P	-0.758	0.05	0	0
2PA	-0.611	0.03	0	0
2P%	0.045	-0.004	0	0
EFG%	0.376	-0.047	-0.017	0
FT	-0.838	0.046	0.143	0
FTA	-0.284	0.031	0	0
FT%	0.108	0.017	0	0.039
ORB	-0.219	-0.049	0	-0.062
DRB	0.104	0.035	0	0.137
TRB	0.353	0.016	0	0
AST	0.761	0.092	0.124	0.427
STL	0.043	-0.014	0	0
BLK	0.206	0.053	0.0225	0.111
TOV	-0.024	0.046	0	0
PF	-0.116	-0.056	0	-0.08
PTS	4.799	0.064	0.189	0.237

According to Table 3 and Table 4, it can be clearly seen that the coefficients of some variables are returned to 0 by Lasso regression and Elastic net regression. Among these four methods, the coefficient of variables related to three-point shooting ability (3P 3PA 3P%) are more significant, there are also some variables related to defense stand out among the whole variables.

This paper calculates RMSE values in the Table 5, the Elastic Net gets the lowest RMSE values, which is followed by OLS regression. Among all variables in OLS regression, VIF values of most variables are greater than 10, indicating that there is multicollinearity between variables. So the Elastic Net model has a better fitting effect. Elastic Net has a smaller RMSE because it combines the characteristics of Ridge and Lasso, i.e., L1 and L2 regularized terms. This allows the Elastic Net model to flexibly adjust parameters with the advantages of both.

Table 4. Regression coefficient (part 2: higher order data)

Variables	OLS	Ridge	Lasso	Elastic Net
PER1	-0.565	-0.004	0	-0.064
TS%	-0.674	-0.057	0	-0.197
3PAr	-0.078	0.012	0	0
FTr	0.203	0.036	0	0
ORB%	0.577	-0.022	0	0.064
DRB%	0.635	-0.006	0	0
TRB%	-1.237	-0.007	0	-0.07
AST%	-0.531	-0.019	0	0
STL%	-0.024	-0.033	-0.001	-0.278
BLK%	-0.059	0.017	0	-0.046
TOV%	0.093	0.042	0	-0.011
USG%	0.206	-0.003	0	0.067
OWS	1.173	0.126	0.024	0
DWS	0.297	-0.044	0	0.249
WS	-1.115	0.075	0	-0.062
WS/48	0.347	0.048	0	0
OBPM	-0.712	0.029	0	0.098
DBPM	-0.284	0.058	0	0
BPM	1.106	0.053	0	0.103
VORP	-0.196	0.123	0.334	0.056

Table 5. RSME

	OLS	Ridge	Lasso	Elastic Net
RMSE	0.510	0.511	0.533	0.507

4. Conclusion

The objective of this study is to investigate the correlation between basketball players' on-court performance metrics and their salary compensation in order to mitigate financial loss for the team's investors. Using data from the 2020-2021 NBA season, a multiple regression analysis was conducted with the combination of basic and advanced metrics as independent variables and player salary as the dependent variable. Descriptive statistics and OLS regression were performed initially, revealing the issue of multicollinearity among the independent variables. Hence, Ridge, Lasso, and Elastic Net regressions were employed to address this issue and screen the 47 independent variables. The model's effectiveness was evaluated using the RMSE method with a validation set, and the results indicated that the Elastic Net method produced the lowest RMSE value, followed by OLS regression. The coefficients generated by the Elastic Net method revealed that investors place higher significance on players' three-point shooting proficiency, defensive abilities, and teamwork capacity, which aligns with the current playing style of the NBA.

References

[1] Stewart B , Smith A. The Special Features of Sport. Annals of Leisure Research, 1999, 2(1): 87-99.

[2] Leeds M A , Kowalewski S. Winner Take All in the NFL: The Effect of the Salary Cap and Free Agency on the Compensation of Skill Position Players. Journal of Sports Economics, 2001, 2(3): 244-256.

[3] Chen Shen. Model Prediction of Factors Influencing NBA Players' Salaries Based on Multiple Linear Regression. Business, 2016, 26: 1.

[4] Huyghe T , Alcaraz P E , Gonzalez J C , et al. The Underpinning Factors of NBA Game-Play Performance: A Systematic Review (2001-2020). The Physician and Sportsmedicine, 2021.

[5] Ying Wang. Exploring The Relationship between NBA Players' Salary and Athletic Contribution. Fujian Sports Technology, 2021.

[6] Tongsi Chen, Chen Cheng, Yujie Chai. Correlation Study between Salary and Team Performance and Player Performance. Sport, 2015, (19): 33-35.

[7] Hakes J K, Turner C. Pay, productivity and aging in Major League Baseball. Journal of Productivity Analysis, 2011, 35(1): 61-74.

[8] Hanhui Wang. Statistical Analysis of NBA Players' Salaries. The World of Survey and Research, 2018, (3): 46-51.

[9] Guodong Tao, Li Cai. Salary Study of NBA Player. Sports Culture Guide, 2012, (4): 5.

[10] Ye Liu. The Effect of Basketball Players' In-game Performance on Players' Salary Earnings - An Empirical Analysis Based on NBA Players' Data. Economic and Social Observation, 2019, (18): 253-254.

[11] Ranstam J, Cook J A. LASSO regression. British Journal of Surgery, 2018, 105.