A Survey of Forecasting the Salary for Players in National Basketball Association

Xingsi Chen [1, †], Haoyang Lin [2, †], Louyang Pang [3, *, †], Yanxiang Wang [4, †]

[1] Seventh High School Zhengzhou, Henan, China

[2] WLSA Shanghai Academy Shanghai, China

[3] The Affiliated HS to Hangzhou Normal University Hangzhou, Zhejiang, China

[4] Xi'an Jiaotong-liverpool University Suzhou, Jiangsu, China

* Corresponding Author Email: yanxiang.wang19@student.xjtlu.edu.cn

[†]These authors contributed equally.

Abstract. The National Basketball Association is the men's professional basketball league in North America. The National Basketball Association league has some of the highest paid professional athletes in the world. The National Basketball Association is the oldest and most prestigious basketball league in the world, which makes the study of the NBA particularly important. During the 2011-2012 season, the season was cancelled after senior staff and team members failed to reach an agreement due to serious differences of opinion in labour management negotiations. The main reason was that the NBA believed that the players' salaries were too high, which left little profit for the NBA and even increased the NBA's losses. Since the ideal interests of management and the interests of labor did not exactly match, both sides made concessions and reached an agreement. Finally, the season resumed. For most NBA clubs, the salary of each player is usually confirmed before the performance in the next game. However, this makes the question of how to determine the amount more difficult. Players' salaries should match their performance. If the player does not meet salary expectations, the club is losing money. If a player finds that he gets paid less than his income based on 61 games that year, the club may lose that player and reputation. In this paper, we summarize previous approaches for salary forecasting and categorize them into: (i) linear regression, (ii) modelling approach, and (iii) non-linear regression and (iv) K-nearest neighbour classifier and provide guiding idea for future work.

Keywords: Predict; Linear regression; KNN; Salary.

1. Introduction

National Basketball Association (NBA) is one of the most commercially successful sports events as well as one of the highest-paid leagues in the world. From 2017-18 season, Stephen Curry earned more than 34.7 million dollars and Lebron James earned more than 33.3 million dollars on the court. Because of the phenomenon of unreasonable salaries that left little profit for the NBA and even increased the NBA's losses, the league has imposed "Salary cap" and "Luxury tax" to restrict both the salaries of players and the expenditure of bosses of teams. Since each team has a salary cap, team executives should be more effective in determining player salaries. However, NBA is evolving, and big data that tracked by technology is contributing to reasonable predictions of player salaries.

Since the beginning of this century, there has been an increasing number of papers on forecasting NBA players' salaries, and the methods are diverse. In 2015, Chen Shen et al. [1] used multiple linear regression to establish the relationship between players' salaries and five variables, which are comprehensive capability, age, attendance rate and winning percentage of team. They also applied stepwise regression method to select variables and utilized Cochran Orcutt's iterative method to eliminate autocorrelation. In 2015, a method [1] that measured a player's quality per minute called Player Efficiency Rating (PER) was created by John Hollinger et al., and they regarded offensive contribution as the most important factor when forecasting the salaries of players. Meanwhile, David Berri, Martin Schmidt, and Stacey Brook et al. [2] created the Wins Produced model (WP) by

researching the relationship between wins and players' efficiency measures. However, they used both offensive and defensive contribution. In 2016, Josh Rosen et al. [3] used variety data such as roster competition, defensive contributions, and impact of bench players by Player Impact Estimate (PIE) to predict future performance to forecast salaries with average positional production. In 2020, Ioanna Papadaki and Michail Tsagris et al. [4] propose a non-linear Random Forest machine learning algorithm (RF) and Least Absolute Shrinkage and Selection Operator (LASSO), a regression analysis method to predict salaries. In this paper, we summarize previous approaches that predict NBA player salaries and divides them into three categories: (i) linear regression, (ii) modelling approach, and (iii) non-linear regression and provide conclusion for our own guiding idea in future work. The remainder of this work is structured as follows. Section 2 describes the methods used in each paper, and summaries, compares, and analyses their strengths and weakness. Section 3 provides a conclusion and our own guiding idea.

2. Main Body

Table 1. Ideas, advantages and disadvantages of the method for NBA player salary forecasting.

Method	Key idea	Strength	Weakness
Linear regression	Only one Factor	Fast and convenient	Too simple
Non-linear regression	Used for abstract models	Probability can be calculated	Too complicated
K-Nearest Neighbor (KNN)	Classification algorithm	High accuracy	Large amount of calculation
Modeling approach	Build a clear model	Make modeling clearer	Methods are too old

Table 1 mainly introduces four methods, namely linear regression, nonlinear regression, KNN, and Modelling approach. Firstly, we introduced their main ideas and advantages and disadvantages. The key idea of linear regression is that there is only one factor, and its advantages and disadvantages are convenience but too simple. The second nonlinear regression can be used for abstract graphics. The advantages and disadvantages are that it can calculate the probability but is too complicated. The third KNN is a classification algorithm. The advantage is that the accuracy is very high, but the disadvantage is that the amount of calculation is too large. The last one is the modelling approach. Their key idea is to establish clear modelling. The advantage is to make the results clearer, but the disadvantage is that the method is too old.

2.1. Linear regression

Table 2. Factor, data, and key idea for publications using linear regression method.

Publications	Factor	Data	Key Idea
LGM [5]	Y	N	Insufficiency of the regression model
Percentage [6]	Y	Y	Forecast percentage, range of difference
MM-estimator [7]	Y	N	Modal linear regression
Statistical Modeling [8]	N	Y	Model estimation, statistical inference, variable

As shown in Table 2, we list the key idea of linear regression method and compare them by factor and data. Linear regression is more common and very suitable for predictive analysis. Since we need to predict the future salary, we can use this method to achieve this goal. We can observe that it uses a linear regression model to predict. This is a better choice, because once the model is created, it can calculate the future salary of all players [5]. The second is to use percentages to calculate this salary. They first use a combination of percentages and linear regression and then calculate. But there will be a problem here, it is difficult to match the percentage and the unit of salary, which will lead to

unclear data. The result will definitely be wrong [6]. The third estimation method is introduced into linear regression. When this model is established, the results obtained will be more concise and clearer. But it is difficult to apply this to the real market, because what we players want is the most accurate number, not an estimate.[7] The last one is to use a statistical model to complete. Of these four, this one is the best to show the most accurate results. Because it not only uses linear regression to predict salary, but also used statistical models to make the result more accurate [8].

2.2. Non-linear regression

Table 3. Factor, data, and key idea for publications using non-linear regression method.

Publications	Factor	Data	Key Idea
Adaptive estimability [9]	Y	N	Estimability presents new discoveries
MLA [10]	N	Y	Importance sampling adaptive
NIRS [2]	Y	Y	Perform-multivariate calibration
NLRM [11]	Y	N	Emphasize estimation and hypothesis testing theoryad

As shown in Table 3, we list the key idea of non-linear regression method and compare them by factor and data. Nonlinear regression is also a more classic method. But it is relatively complicated. First, the first one is to use adaptive estimability to predict. The advantages and disadvantages of this method are similar to those of the previous estimation method, which puts simplicity and speed in the first place, ignoring accuracy [9]. The second is to use a machine learning method to predict salary. This idea is very good. It also combines Bayesian theory and nonlinear regression. The result analysis is also better [10]. The third is to create a model through an example of NIRS, and then use this model to predict salary. The disadvantage is that it is too complicated [2]. The last one is to use a non-linear model to predict. This is a relatively modest method, which is not very accurate but the result is acceptable. It can be used as an experiment. I think the second of these four is relatively the best [11], since machine learning is a comprehensive approach to predict data. Ioanna Papadaki and Michail Tsagris, et al. propose a non-linear Random Forest machine learning algorithm (RF) that could avoid over-fitting of linear regression method. They first standardize the data and then apply sophisticated models. RF algorithm is a data mining method that is fast and flexible for high dimensional data. It selects variables and bootstrap samples randomly and repeated and aggregate the results: creating many trees and compound a forest. The advantage of RF and variable selection algorithms is their ability to capture non-linear components. Least Absolute Shrinkage and Selection Operator (LASSO) is a regression analysis method that progress variable selection and regularization of the regression coefficients. This method can enhance the accuracy and interpretability of regression models as well as making the model more interpretable. LASSO algorithm is helpful when detect the most important performance measures [4].

2.3. K-Nearest Neighbor (KNN)

Table 4. Factor, data, and key idea for publications using K-nearest neighbour method.

Publications	Factor	Data	Key Idea
Pattern classification [12]	Y	Y	Improved kNN algorithm
ML-KNN [13]	Y	Y	Label set of adjacent instances
RSS [3]	Y	Y	Data mart
Intrusion detection [14]	N	Y	The KNN classifier can effectively detect intrusive attacks

As shown in Table 4, we list the key idea of K-nearest neighbor method and compare them by factor and data. The power of K-nearest neighbour classifier method lies in classification, and its algorithm is also very powerful, so that its disadvantage is that the amount of calculation is too large and the whole process is relatively complicated. The first one is to use one of the model classifications in this method to classify the factors related to salary, and then to predict the final salary. The accuracy

of this method is very good, taking into account many factors [12]. The second one is ML-KNN. This is an upgrade of the original KNN. The concept is that each factor will have its own label, and then this label will show the size of their impact on salary. Doing so makes the analysis more convenient and the results clearer. The case is that in bioinformatics, a protein may have many effects on cells when predicting its function class. [13]. The third is to use RSS combined with KNN to make predictions. RSS can analyze data very well, so it is very useful for predicting the salary of national association basketball players [3]. The last one uses an intrusion detection model to predict player salaries, coupled with the KNN algorithm, which is a relatively novel method. There is another problem, too complicated and too much calculation [14]. In Shi Cheng et al.'s Opta Player Big Data Prediction: NBA Player Data Analysis and Salary Forecasting, he looked for the value of k in the process of model tuning after importing the data. Then he applied KNN algorithm in both small-size data and big-size data and created means and standard deviations of training set scores and validation set scores and drew a plot. He set a KNN classifier with k neighbours when k is equal to 5 and prediction accuracy is 0625. This model weighs standard deviation and variance, thus makes model tuning easier [1]. Hence, every method can effectively predict salary except the last one.

2.4. Modeling Approach

Table 5. Factor, data, and key idea for publications using modelling approach.

Publications	Factor	Data	Key Idea
Information retrieval [15]	N	Y	Probabilistic language modeling
MMA [16]	Y	Y	Improvement of modeling approach
SEM [17]	Y	Y	Support model for multiple path assumptions
SMA [18]	N	Y	Statistical signal power model

As shown in Table 5, we list the key idea of modelling approach and compare them by factor and data. Player Efficiency Rating (PER) was created by John Hollinger to measure a player's quality per minute, which was the combination of a player's contribution based on player's offensive statistics. David Berri, Martin Schmidt, and Stacey Brook et al. created the Wins Produced model (WP) by researching the relationship between wins and players' efficiency measures in the book The Wages of Wins. Calculating WP48 and Wins Produced needs the "offensive efficiency" and "defensive efficiency". Andrew Thomas Fleenor et al. created a model by using individual statistics that contains raw numbers and per game averages in their article Predicting National Basketball Association (NBA) Player Salaries. They identified which factor related most to salary via a dummy variable field. This method improves the accuracy of the results because extremely large or small data are discarded in the model to obtain more general results. The method of Hollinger focuses too much on players' offensive efficiency instead of their defensive efficiency. Berri's model, however, does not show the total overall contribution for the team of a player. His model estimates the value of a game offensive and defensive statistic. Andrew Thomas et al. propose a method which is reasonable due to the extremum is discarded, weaken the effect of special values on the results, such as LeBron James. Modelling approach is a relatively old method, and thus in the process of analysing data and forecasting, it will be inferior to the first three, but his result analysis is better. The first one is to combine Information Retrieval to help create better models, because data filtering is very important. His method just made up for the shortcomings of the modelling approach [15]. The second is MMA,[16] which is also a mathematical model, which is also helpful for data analysis. So one and two are relatively similar. The difference is that the second method directly uses the model to analyse the data, which is more convenient and faster [16]. The third, SEM model, can help select factors to predict salary [17]. The last one is the statistical modelling method, SMA. It is also a model that can better help predict salary [18]. Therefore, the second model can achieve the goal of predicting salary with high accuracy.

3. Conclusion

In this paper, we mainly describe four methods to predict the future salary of NBA players. They are linear regression, nonlinear regression, KNN algorithm, and modelling approach. First, linear regression, this is what we all commonly use, it is very convenient and easy to understand,[5] but the shortcomings are also obvious, there can only be one factor. So, the accuracy will be weaker. Nevertheless, it has some wide usages in our daily life, especially in predicting of population of animals. It can help a lot of environmental science. On the contrary, nonlinear regression is better. This method is not restricted to a straight line, so it can express more abstract models, which can be applied to various problems in life. To be more specific, climate change can be predicted in a small area using this method. Secondly, the KNN algorithm is a very good classification method, which plays a very important role in the selection and prediction of factors such as selecting variables from weather and predicting how much rainfall is crops need. But KNN's drawback is that the amount of calculation is very large, and the requirements for data sorting are very strict. If the sample data differs too much, the accuracy of the results will be greatly reduced. The last one is the Model approach. This model provides a great help to our forecasted salary. He builds a model to clearly express the forecast results. When this model is perfectly completed. For the entire market, not just the NBA, will make a huge contribution. This is very helpful for the management of each team. This model allows management to use the team's salary more reasonably and avoid exceeding the team's total salary. It can also avoid giving players a salary that is too high or too low. Because the operation of the entire team has been greatly improved. The future direction of development is to use forecasted wages as a bridge and continue to expand to various fields, such as wage forecasts in other leagues, such as football and volleyball.

References

[1] Shen Chen. (2016). Research on the Influencing Factors of Nba Players' Salaries. Shang (26), 46.

[2] Perez-Marin, D., Garrido-Varo, A., & Guerrero, J. E. (2007). Non-linear regression methods in NIRS quantitative analysis. Talanta, 72(1), 28-42.

[3] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90-108.

[4] Papadaki, I., & Tsagris, M. (2020). Estimating NBA players salary share according to their performance on court: A machine learning approach. arXiv preprint arXiv:2007.14694.

[5] Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. Transactions of the Institute of British Geographers, 145-158.

[6] Zhao, L., Chen, Y., & Schaffner, D. W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. Applied and environmental microbiology, 67(5), 2129-2135.

[7] Yao, W., & Li, L. (2014). A new regression model: modal linear regression. Scandinavian Journal of Statistics, 41(3), 656-671.

[8] Groß, J. (2012). Linear regression (Vol. 175). Springer Science & Business Media.

[9] Manski, C. F. (1984). Adaptive estimation of non–linear regression models. Econometric reviews, 3(2), 145-194.

[10] Blum, M. G., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. Statistics and computing, 20(1), 63-73.

[11] Takeshi Amemiya, Chapter 6 Non-linear regression models, Handbook of Econometrics, Elsevier, Volume 1, 1983, Pages 333-389.

[12] Sun, S., & Huang, R. (2010, August). An adaptive k-nearest neighbor algorithm. In 2010 seventh international conference on fuzzy systems and knowledge discovery (Vol. 1, pp. 91-94). IEEE.

[13] Zhang, M. L., & Zhou, Z. H. (2005, July). A k-nearest neighbor-based algorithm for multi-label classification. In 2005 IEEE international conference on granular computing (Vol. 2, pp. 718-721). IEEE.

[14] Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. Computers & security, 21(5), 439-448.

[15] Ponte, J. M., & Croft, W. B. (1998, August). A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 275-281).

[16] Sharomi, O., & Gumel, A. B. (2008). Curtailing smoking dynamics: a mathematical modeling approach. Applied Mathematics and Computation, 195(2), 475-499.

[17] Gursoy, D., Jurowski, C., & Uysal, M. (2002). Resident attitudes: A structural modeling approach. Annals of tourism research, 29(1), 79-105.

[18] Roos, T., Myllymaki, P., & Tirri, H. (2002). A statistical modeling approach to location estimation. IEEE Transactions on Mobile computing, 1(1), 59-69.