Predicting NBA Player’s Salary Based on Statistics from the Game Using Linear Regression

Mohan Cao
School of Jinan University and Birmingham University Joint Institute, Guangzhou, China

Keywords: Machine Learning, Linear Regression, Prediction, NBA Players, Statistics.

Abstract: For professional sport league like NBA, settling a proper salary for players is vital for the long-term development of a team, since hold players whose performance is out of proportion with its salary could harm the team, especially in a league that set the upper bound of salary. However, it’s hard to quantify a player’s performance by data, then use the “performance” to further predict the salary. Thus, this study focuses on predicting NBA players’ salary by using linear regression of multiple factors to find a direct relationship between specific kind of data on court that correlates to salary at a high extent. The main procedures of this study contain the characterization of different data, which determine whether they correlate well with salary; the processing of the rough data; and the application of the linear regression model with multiple variables that have a strong relationship with the salary. The result of this program is a model that could predict the salary of player in the next few coming season. With this model, teams and fans could predict the salary, given the dataset of those players. The original salary could be a reference to determine the accuracy of the prediction. Also, the weightings and factors of formula in this work could be slightly changed to meet the demands of different cases.

1 INTRODUCTION

There exist multiple kinds of statistic that could measure a player’s contribution to their team and quantify their performance on the court, including points, rebounds, assist, turnovers and other data as well as its corresponding salary. These statistics measure players contribution from different perspective (Shah and Romijnders 2016). For instance, average points per game could quantify the ability of players to score and assists could demonstrate a player’s ability in creating opportunities for teammates to score. Different categories of data demonstrate different kinds of contribution to the basketball game, whose combination in turn could roughly reflect the performance of a player.

With the purpose of sustaining the balance between the basketball teams in the league, salary cap, which means a limit for the overall sum of salary, had been implemented for all the teams in NBA (Kesenne and Scott 2000). Considering the salary cap, decision makers of basketball teams should determine the salaries of basketball players carefully (Nagarajan et al 2018). Therefore, executive officer in this league wish to find a model to predict the salary and then give contract with salary which accord with a prediction model.

However, it’s complicated and, in some cases, not accurate to measure the performance of a player using those data on court directly. Different positions of player may focus on making contributions to various aspects of the team, for instance, guard players may have more assists, while centre players and power forward will take more responsibility on defence and rebounds (Vitor et al 2018). As a consequence, it’s, in some cases, biased to measure the performance of players by statistics. Additionally, it’s difficult to classify players to different categories, then construct several models to measure performance by the statistics for a specific category of players. The development of the basketball and the modern NBA causes the blend on duties for different positions. For instance, in modern professional basketball games, some centre players regularly carry out the duty of point guards and shooting guards, vice versa (Su et al 2012). Also, it is common for players to play multiple

Cao, M.
Predicting NBA Player’s Salary Based on Statistics from the Game Using Linear Regression.
DOI: 10.5220/0012823900004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 476-480
ISBN: 978-989-758-690-3

positions on court and switch their positions based on the need of their teams or the situation on court.

So instead of constructing a model to measure the "overall performance" of players by statistics, then using performance to predict the salary, this study focuses on finding the relationship between each kind of data (involving points, assists, etc.) and the corresponding salary directly, if the relationships exist and are strong enough, then using this potentially exist relationships to predict the salary. For example, assume that points per game have a direct mathematical relationship to salary, then the goal is to find the relationship between points per game with salary (Gareth et al 2021). After finding several similar relationships regarding of those statistics and salary, we could construct a model to predict the salary, given the statistics of each player. The initial step of this study is to analyst statistics and find data that could potentially hold high correlation to salary. After searching into the relationship of those categories of data and the salary, a model that could predict salary based on given data can be constructed. Then, using this model, we can get the result of e salary prediction by simply input the original data.

2 METHODS

2.1 Data Source

Captions should be typed in 9-point Times. They should be centred above the tables and flush left beneath the figures. The data set that involves various data and salary is collected from www.kaggle.com. This dataset involves the average points, assists, blocks, age, turnovers, field goals rates of different kinds of shooting, which involves 2 points shot, 3 points shot and free throws per game of every player in the NBA in season 2022-2023. Over a relatively long period, the level of player could be unpredictable. Also, the salary cap experienced serious fluctuation in recent decade and is predicted to increase in the near future. As a consequent, the result for this model may only be valid in the following 2 or 3 seasons.

2.2 Indicator Selection and Description

The data of field goals rates are omitted since this data couldn't measure the performance of players precisely. In most cases, centres have a higher field goals rate than guards and forward as most of their shooting are closer to the basket. Also, star players will bear more defensive pressure, which could lower their overall field goals rate. In contrary, role players get more opportunities for wide open shots. Other data are reserved (Morgulev et al 2018).

2.3 Introduction of Methods

Linear regression plays a fundamental role in constructing statistical model, which could be applied in prediction a quantitative amount (Gareth et al 2003). Linear regression of single/ multiple factor(s) is a linear approach for modelling a predictive relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables), which are measured without error. Consequently, this model meets our demands. The initial step of this research is to find data which potentially have a high correlation to salary (Papadaki and Tsagris 2020). By constructing scatter diagrams, we could find specific data that seem to have a strong linear relationship to salary. Since considering all reserved data simultaneously is complicated and could lead to error, this step aims at the one-to-one correspondence relations between data and salary. The next step is to find the correlation coefficient of specific data to confirm previous result and to determine specific statistics that will be input into the model of linear regression. The final step is using the model of linear regression to find a linear relationship between specific data and the result: the predicted salary. This is an application of machine learning for quantitative and qualitative predictors (Emirman et al 2021). Following this method, a function regarding of specific data and salary could be found.

3 RESULT AND DISCUSSION

3.1 Data Visualization

The scatter diagram graphs pairs of numerical data, with one variable on each axis, to look for a relationship between each other. It could show a visual relationship among each statistic and their corresponding salary. Following this method, we can pick out data that potentially have a high correlation with salary.

As shown in figure 1, 2, 3 points, assists and turnovers potentially have a positive linear relationship with salary.

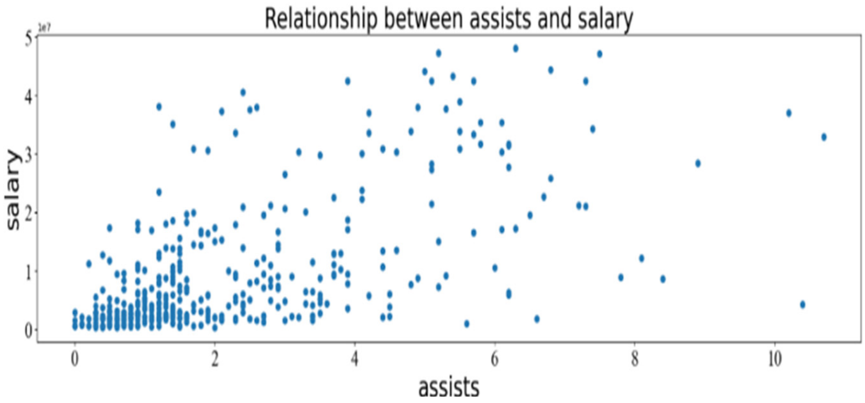

Figure 1: Relationship between assists and salary (Picture credit: Original)

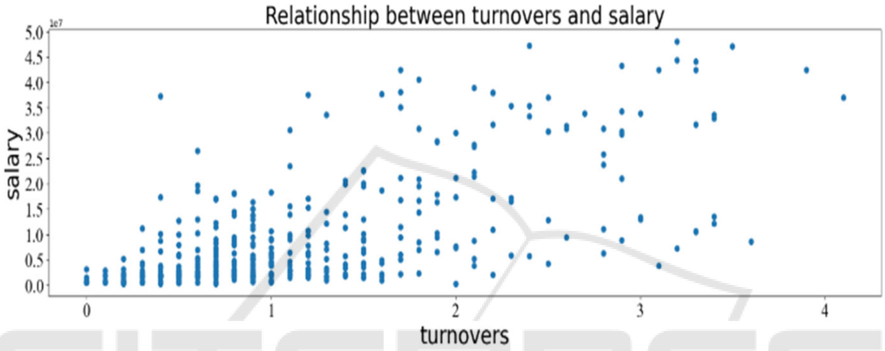

Figure 2: Relationship between turnovers and salary (Picture credit: Original)

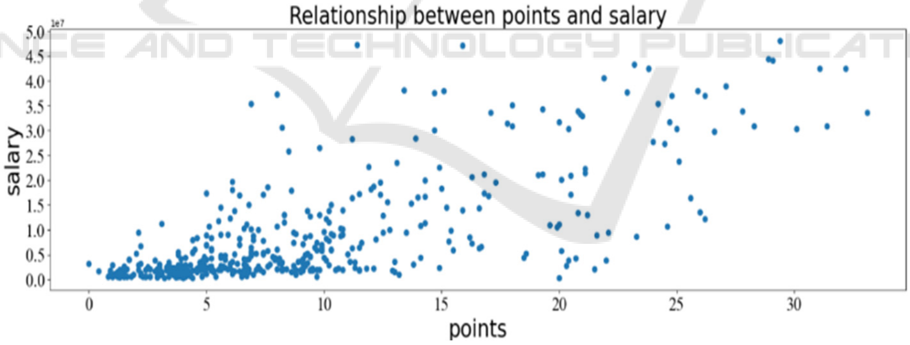

Figure 3: Relationship between points and salary (Picture credit: Original)

3.2 Data Processing

As shown in table 1, the average points, turnovers and assists have a relatively high correlation coefficient, with 0.728, 0.670 and 0.626, respectively, when compared with other data (all under 0.50). The interpretation of the result is as follows: points and assists evaluate players' contribution to score, which is the most essential and fundamental skill for a player.

Table 1: Correlation between data and salary.

Data	Correlation	Relationship
Points	0.728	Strong
Assists	0.626	Strong
Turnovers	0.670	Strong
Blocks	0.296	Weak
Ages	0.413	Weak
Rebounds	0.475	Weak
Steals	0.489	Weak

The number of turnovers evaluates the mistakes players made on court, which has a strong relationship to salary. Personal interpretation to this is that more mistakes made is on behalf of more attempts in decisions-making on court, which demonstrate their importance as well as irreplaceability in their team.

3.3 Application of Linear Regression Model

Based on previous result, points, turnovers and assists should be input into the multiple linear regression model. This model constructs a linear relation regarding the three data and the corresponding salary prediction. And according to the model, the final result is:

$$\boldsymbol{y = 946567 * ['PTS'] + 1149793 * ['AST'] - 478153 * ['TOV'] - 1858221} \quad (1)$$

3.4 Discussion

The model has met the initial demand of this model, that is offering a valid prediction of salary based on the data of players. The advantage of this method is that the rough data which is relatively accessible is all the linear model needed. Additionally, this model can deliver a prediction of all categories of player. Though linear relationship of diverse data and the corresponding salary prediction is figured out, this model still has its systematically inevitable limitations. The dataset originated from the statistics in season 2022-2023, and this model could be invalid after several seasons with the burgeoning increasement of salary map in future. Due to these reasons, the gap between prediction result from this method will grow It is noted that this model hasn't considered regarding the combination of two data as a new data itself, which could hold a potential stronger correlation relationship with the salary, because of the complexity of permutation among those data. For example, average points multiply the field goals rate may be more comprehensive data that reflects the ability to score and rebounds multiply the frequency to block could reflect a player's ability to protect the paint area comprehensively, thus the combination potentially have a stronger relationship with salary. Also, to optimize the overall prediction result, data for specific players whose salary is obviously out of portion with its performance could be omitted. The common case for disproportion among data and salary is injuries and rookie contacts for extremely talented young players. Anyhow, the residual error and the overall degree of fitting is acceptable; also, the final result constructed by this model accords with intuition in most cases.

Turnovers itself have a positive linear relationship to the salary, but when it comes to taking consideration of multiple factors together, turnovers have negative impact on the predicted salary. The personal interpretation to this phenomenon is as follows. More turnovers mean more wrong decisions on court, revealing that a player makes more decisions on court, which demonstrate their significance to the team, as the author has mentioned before. Such that, when considering the turnovers as a single factor to salary, it has a positive impact on salary. However, in the case that we consider multiple factors, points and assists can also measure a player's significance for the team, as they mean more shots made and contribution in ball movements, respectively. When the "significance" of players is similar, that is to say, they take on the same portion of responsibility on the offensive side on court. Thus, less turnovers prevail that their overall decisions made are more accurate and efficient, such that in this case, turnovers have a negative impact on the performance as well as the income prediction of the player. Additionally, as mentioned before, salary of players which experience serious injuries, rookie players and top superstars in the league could not be measured by this model. The reason is that in those cases either data couldn't reveal the true value or contribution of those players or those players have unique commercial values for their teams that couldn't be measured by data on court.

4 CONCLUSION

This model meets the initial aim of presenting a direct relationship between data for basketball players and their corresponding salary in national basketball association. As a consequence, this enable the prediction of the amount of their contract amount, as long as their statistics is given. The main finding of this study is finding a function regarding of points, turnovers, assists and salary. More specifically, each point obtained increases the salary by 946567 dollars per year; each assist made increases the salary by 1149793 dollars per year; in contrary, each turnover decreases the amount of salary by 478153 dollars. The final prediction is the sum of previous three functions subtracted by constant 1858221, according to the linear regression model. Take Seth Curry-a guard player in Maverick as an example. In season 2022-2023, the player got 9.2 points, 1.6 assists and 0.8 turnover per game, and the predicting salary is

8023372 dollars. This formula emphasizes the importance of scoring as data that correlates with scoring determine the final result in a high extent, because the number of points made is way higher than the other data, despite their discrepancy in coefficient. Actually, the salary for Seth Curry in season 2022-2023 is about 8.5 million dollars, slightly higher than estimated salary. According to the model, the player deserved slightly less money in the following 2023-2024 season. The overall finding is closely related to our initial thought as it meets the demand of offering a straight-forward way to predict salary regardless of measuring the overall "performance" of player. Fans of the NBA league as well as professional players could consider this model as a reference to their estimated salary, thus have psychological expectation. It is noted that some factors other than on court statistics can also affected the estimated salary, such as ages, injuries experience, potential and years played for a single team, which tend to increase their overall commercial value and guarantee the tacit agreement among their teammates. However, executive directors in each team could follow this method to measure the rough value of a player to compose the best NBA team. Also, weighting different values of data could potentially meet their demands more accurately in multiple cases. This model is just a foundation for the rise of advanced analysis that could enabled more accurate empirical valuations of players. Also, the value of this finding also lies in its function to measure the predicted salary of a player directly for anyone that is interested in the domain of professional basketball.

REFERENCES

R. C. Shah, R. Romijnders, Appl. Bask. Traj., (2016)

S. Kesenne, J. Scott, Polit. Econ., 47(4), 422–430 (2000)

R. Nagarajan, Y. Zhao, L. Li.: Inter. Conf. Big Data Ana., 138–143 (2018)

C. Vitor, J. I. Sérgio, L. G. N. Eduardo, F. B. Adriano, Jua. Viei. Nasc., 23, (2018)

X. Su, X. Yan, C. L. Tsai, Lin. reg., (2012)

J. Gareth, W. Daniela, H. Trevor, T. Robert, STS, (2021)

E. Morgulev, O. H. Azar, R. Lidor, Int. J. Data Sci. Anal. 5, 213–222 (2018)

J. Gareth, W. Daniela, H. Trevor, T. Robert, Spring. Sci. Bus.Med., (2003)

I. Papadaki, M. Tsagris, Mac. Learn. Appr., (2020)

O. Emirman, Y. Mucahit, K. Tolga, Lec. Notes Net. Sys., (2021)