National Basketball Association Salary Prediction: A Data-Driven Linear Regression Analysis

Jiahe Zhang *

Department of Math, University of Birmingham, Birmingham, The United Kingdom

* Corresponding Author Email: jxz184@student.bham.ac.uk

Abstract. National Basketball Association (also known as NBA which will be used as an abbreviation throughout this essay) as one of the few most successful professional sports leagues in the world, it has been well known by the fierce physical competition among the most talented and competitive players in the realm of basketball. However, there is something else to discover behind the dazzling crossovers and sensational clutches, hidden in the reflection on the O 'Brien Cup, for example, the salary. This essay will be discussing the prediction on NBA players' salary using multiple supervised machine learning algorithms based on a dataset of players' on-court data and achievements, aiming for an objective result that can be used for both prediction and evaluation of players' contracts. The result of this essay reveals which specific variables are useful for predicting players' salary.

Keywords: National Basketball Association; player salary prediction; supervised machine learning.

1. Introduction

In NBA, the best way to make profit has been to win the championship. Naturally, by recruiting or trading valuable players to build a championship-winning team is the goal for every club. To make the competition fair, there is a maximum amount each team is allowed to invest on players every season, which is so called 'salary cap', if such limit was reached, luxury tax is required to pay towards the association. Therefore, a manageable budget and spending the money efficiently are crucial. This is where a prediction algorithm should be coming in. Not only to manage the money flow, but also to evaluate the values of current players' contracts in the team so that the decision-makers can construct the team better.

Based on the research objective on exploring the salary, relevant data was collected and comprehensively researched by advanced machine learning method. The results in this paper show that Championship factors, Playoffs, Game Score in regular seasons, Game Score in playoffs, Win Share in playoffs, Game Score in regular seasons, Win Share in playoffs, All-time leading records in finals. The structure of this essay is as follows: Section 2 shows some related literature review, Section 3 illustrates the data, methodologies used in this paper as well as the results, and Conclusion is in Section 4.

2. Literature Review

Data and statistics have been heavily influencing every sport in this era, which can objectively enhance decision makers' work for a grand amount. Sports analytics is an emerging field that grabs data to optimize the decision-making process. With it, teams can have better on-court winning strategies, off-court training plans and other approaches to improve athletes' performance. By the year of 2028, it is expected that the sports analytics industry will profit $3.4 billion globally, as presented in the article by Schroer and Urwin in 2022. There have already been some great papers researching on the factors influencing NBA players' salary. A paper in 2018 by Sigler and Compton claimed that the determining factors of NBA players' pay are experience, points, rebounds, assists and fouls, whereas the number of 3-point shots and Hollinger's player efficiency rating (as more commonly known as PER) are insignificant . On the other hand, Nagarajan and Li (in 2017) conducted research aiming for the best strategy to select players through (1) players' on-court statistics (2) performance of teams (3) salary cap, in which they explained fundamental rules and concepts of NBA

and the method of measuring players' on-court efficiency as well as the connection of efficiency with performance of teams. Salary cap 1)prevents wealthy teams to dominate in the league to balance the competition; (2) avoids overmuch expense on the top players for a fairer salary distribution in a team; (3) ensures both large and small club owners can expect a reasonable profit, as research by KeÂsenne in 2000. However, there were only 450 players in NBA in 2021, it was relatively low (compared with 1,696 players in The National Football League, 780 players in Major League Baseball), which makes it difficult to build a good model. Besides, data from video games of basketball (2K series) is as useful as that from real basketball world. As researched by Kahraman, Cebi, Onar, Oztaysi, Tolga and Sari in 2021, the variables in the video game 2K20 profoundly contributed to predict NBA players in the real world. As a matter of fact, the predictions were very close to the salaries in season 2021-2022. The combat of basketball is nothing more than offense and defence. However, in the research of Baghal in 2012, it seems that NBA players' salary only has direct relationship with players' offense quality not defence quality. Higher investment on salary brings team more winning. Nevertheless, Baghal suggested that more expense should be focusing on improving defence quality. Beyond performance on the court, Papadaki and Tsagris (in 2022) suggested that factors like popularity and spectacle can also affect salary predictions significantly. Ertug and Castellucci also did research (in 2013) on how the two off-court factors: reputation and status of resource providers influence players' salary. The results showed that they both turned out to be positively influencing salary, between which status resource providers has a stronger relationship with salary than reputation with salary. Age of players also seems to affect their salary. Hentilä (in 2019) revealed that rookies (new players) in NBA tend to get underpaid, whereas players are likely to get overpaid as they age.

3. Methodology and Results

3.1. Data Pre-processing

It was researched by Bowen in 2020, an NBA player performance scoring system created by the researcher. There are 17 variables clustered into 4 groups and each variable was assigned with a different weight so that a player's total score can be calculated through this algorithm. The variables are: Accomplishment: (1)personal awards (2)first/second/third teams (3)on-court data (4)Most Valuable Player shares (5)championship factors (6)playoffs; Commitment: (7)Game Score in regular seasons (8)Game Score in playoffs (9)Win Share in regular seasons (10)Win Share in playoffs; Prime: (11)Game Score in regular seasons (12)Game Score in playoffs (13)Win Share in regular seasons (14)Win Share in playoffs; Legacy: (15)all-time leading records in regular (16)all-time leading records in playoffs (17)all-time leading records in finals.

Before building any models, the datasets taken from and need filtration. As the players come from very different periods, the model would not be accurate considering inflation, changes on salary clauses and salary caps. Hence, this essay only studies on the top 50 active players from the year of 2000 to the year of 2022.

Fig. 1 is the Correlation Matrix of all the variables (Players' names were dropped). In this matrix, the lighter the colour of the block is, the higher correlation of the two corresponding variables have. For example, obviously, 'Avg_salary' (Average Salary) has correlation 1.0 with itself, whereas it barely has correlation with 'Legacy_final'.

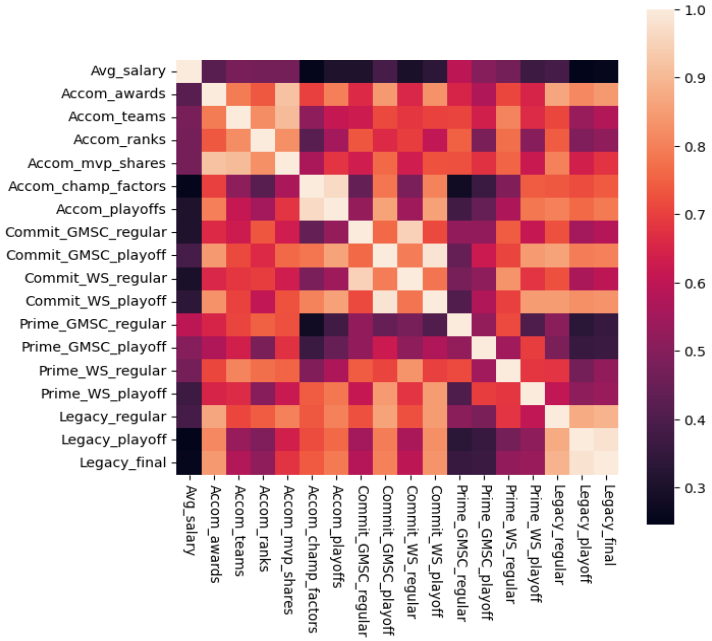

Fig. 1 Correlation Matrix.

3.2. Prediction

There are 2 regression models being used in this essay: Simple Linear Regression as well as Multiple Linear Regression, whose results will be compared at the end of Section 3.

3.2.1. Simple Linear Regression

Simple Linear Regression is a Linear Regression model with only one independent variable as the predictor to predict the dependent variable which always is Average Salary in this essay. The first model starts with Average Salary versus Total Score (a sum of weights of all independent variables). Fig. 2 contains the scattered data, and they are described by a curve, which suggests no direct linear relationship between the independent variable Total Score and the dependent variable Average Salary. In Table 1, they are two results of different ways splitting the dataset. They have Mean Squared Error (an indicator on the deviation of data, as model error increases, MSE increases) of 11.97 and 19.62, respectively. However, both results are not satisfying since the dependent variable, Total Score, is the sum of all the 17 variables from , and some of the variables help to predict Average Salary, but some do not. Hence, Simple Linear Regression does not suffice to be a good model in this case.

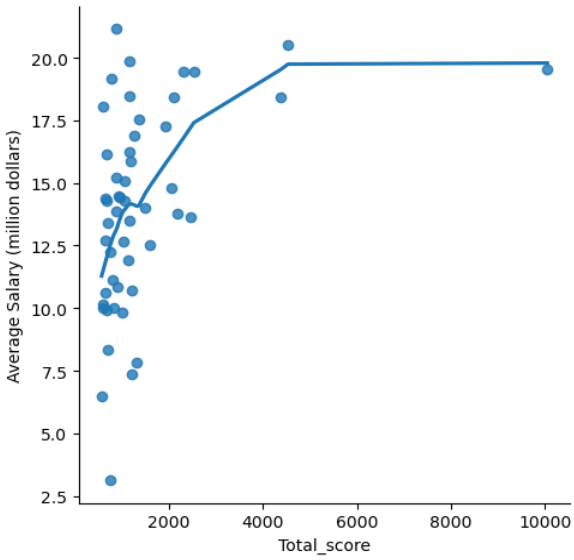

Fig. 2 Simple Linear Regression result.

Table 1. Training results.

	Size of train set	Size of test set	Mean Squared Error
Model 1	40	10	11.97
Model 2	45	5	19.62

3.2.2. Multiple Linear Regression with Backward Selection

Multiple Linear Regression is a prediction model that takes multiple variables as predictors to regress with the dependent variable, which can show which predictors do help explain the dependent variable and which do not. Table 2 was attained as a summary after transforming and fitting Multiple Linear Regression of all 17 variables against Average Salary:

Table 2. Multiple Linear Regression summary.

	Coefficient (β)	Standard Error	t-statistic	p-value
Intercept	-13.51880	10.04800	-1.34500	0.18800
Personal awards	0.00810	0.00900	0.91200	0.36900
First/Second/Third teams	0.01290	0.01500	0.88500	0.38300
On-court data	0.01050	0.01300	0.81800	0.41900
Most Valuable Player shares	-0.05950	0.03300	-1.78600	0.08400
Championship factors	13.38820	9.13500	1.46600	0.15300
Playoffs	-0.09750	0.05400	-1.81900	0.07800
Game Score in regular seasons	-0.11150	0.08900	-1.25600	0.21800
Game Score in playoffs	0.38750	0.13700	2.82100	0.00800
Win Share in regular seasons	0.00880	0.10400	0.08400	0.93300
Win Share in playoffs	-0.31850	0.13200	-2.40600	0.02200
Win Share in playoffs at prime	0.06820	0.03900	1.76100	0.08800
Win Share in playoffs at primes	0.03180	0.03500	0.89600	0.37700
Win Share in regular seasons at prime	0.02700	0.05400	0.50400	0.61800
Win Share in playoffs at prime	0.03900	0.04900	0.79600	0.43200
All-time leading records in regular	0.03600	0.07600	0.47400	0.63900
All-time leading records in playoffs	-0.03470	0.05700	-0.61000	0.54600
All-time leading records in finals	0.16320	0.14700	1.11000	0.27500

In Table 2, coefficient βi refers to the average change on the dependent variable, Average Salary, of one unit increase in predictor Xi while other predictors remaining fixed; t-statistic refers to indicate the linearity of each predictor with Average Salary; p-value indicates the significance of the variable predicting Average Salary. P-value indicates the significance of predictors, where a small p-value suggests the corresponding predictor is significant. Conversely, a large p-value suggests that the corresponding predictor is not significant, i.e., the predictor does not help explain the dependent variable. With this table, and the formula: Average Salary = β0 + β1* 'Personal awards' + β2 * 'First/Second/Third teams' + β3* …. +β17 * 'All-time leading records in finals' + ϵ (random error), a player's predicted salary can be calculated if the values of all the variables and coefficients are known. Nevertheless, as mentioned before, some predictors are not helpful as much as others. For instance, predictor 'Win Share in regular seasons' has a very large p-value of 0.93, which indicates that it barely has significance on predicting Average Salary. Variables like this should be dropped one by one and fitting the model again so that the model can be optimized, such method is so called Backward Selection. The next step was to find the best scenario by trying different times of Backward Selection. The average p-value kept decreasing by doing more times of Backward Selection until the 10th time, where the average p-value then increased. Hence nine times of Backward Selection is the best scenario in this problem and the summary of the result is shown in Table 3 in which the remaining variables are naturally the most significant ones, which are Championship factors, Playoffs, Game Score in regular seasons, Game Score in playoffs, Win Share in playoffs, Game Score in regular seasons, Win Share in playoffs, All-time leading records in finals, to predict the dependent variable – Average Salary.

Table 3. 9 times of Backward Selection.

	Coefficient	Standard Error	t-statistic	p-value
Intercept	-10.92450	8.33100	-1.31100	0.19700
Championship factors	11.95100	7.24600	1.64900	0.10700
Playoffs	-0.08500	0.04000	-2.13400	0.03900
Game Score in regular seasons	-0.06050	0.02500	-2.39000	0.02200
Game Score in playoffs	0.29880	0.09800	3.04000	0.00400
Win Share in playoffs	-0.24030	0.09100	-2.64700	0.01100
Game Score in regular seasons	0.08570	0.02000	4.32200	0.00000
Win Share in playoffs	0.05110	0.03300	1.53300	0.13300
All-time leading records in finals	0.08380	0.05800	1.43400	0.15900

Table 4 shows the average p-values and Mean Squared Errors of all the models in this essay. Average p-value was 0.3238 with all 17 variables being used for Multiple Linear Regression without Backward Selection. After Backward Selection, the average p-value is now 0.0747, which decreased 76.93%. And the Mean Squared Errors of both Multiple Linear Regression models are evidently smaller than that of Simple Linear Regression models.

Table 4. Average p-values and Mean Squared Errors.

	Average p-value	Mean Squared Error
Simple Linear Regression (1)	/	11.97
Simple Linear Regression (2)	/	19.62
Multiple Linear Regression without Backward Selection	0.3238	6.80
Multiple Linear Regression with Backward Selection	0.0747	7.90

4. Conclusion

In conclusion, Simple Linear Regression and Multiple Linear Regression were used to build models to predict NBA players' salary in this essay. Clearly, the performance of the model of Multiple Linear Regression with Backward Selection is much better than that of Simple Linear Regression in this prediction problem for the reason that Backward Selection eliminated the variables that are not valuable for predicting salaries. The effect of doing Backward Selection on the original Multiple Linear Regression model was significantly positive, which lessened p-value of 76.93% but with a slightly larger Mean Squared Error than not doing Backward Selection. The final result reveals that the most significant factors influencing NBA players' salary are Championship factors, Playoffs, Game Score in regular seasons, Game Score in playoffs, Win Share in playoffs, Game Score in regular seasons, Win Share in playoffs, All-time leading records in finals.

However, there are some other factors affecting the results of the models. (1) Rookie player: Some of the players in the dataset had very short experience in NBA, they surely could not have signed big contracts (2) Small size of dataset: as mentioned in literature review, the number of players in NBA is not large and only 50 players were considered in this essay. More players should be considered in the future.

References

[1] Morgulev, E., Azar, O. H., & Lidor, R. Sports analytics and the big-data era. International Journal of Data Science and Analytics, 2018, 5: 213-222.

[2] Sigler, K., & Compton, W. NBA Players' Pay and Performance: What Counts? Sport Journal, 2018.

[3] Nagarajan, R., & Li, L. Optimizing NBA player selection strategies based on salary and statistics analysis. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 2017: 1076-1083.

[4] Késenne, S. The impact of salary caps in professional team sports. Scottish journal of political economy, 2000, 47(4), 422-430.

[5] Kahraman, C., Cebi, S., Onar, S. C., Oztaysi, B., Tolga, A. C., & Sari, I. U. Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, 2021, 2, Springer Nature.

[6] Baghal, T. Are the" four factors" indicators of one factor? An application of structural equation modeling methodology to NBA data in prediction of winning percentage. Journal of Quantitative Analysis in Sports, 2012, 8(1).

[7] Papadaki, I., & Tsagris, M. Are NBA Players' Salaries in Accordance with Their Performance on Court? In Advances in Econometrics, Operational Research, Data Science and Actuarial Studies: Techniques and Theories, 2022: 405-428.

[8] Ertug, G., & Castellucci, F. getting what you need: How reputation and status affect team performance, hiring, and salaries in the NBA. Academy of Management Journal, 2013, 56(2): 407-431.

[9] Hentilä, D. The link between salary and performance: Are NBA players overpaid? Retrieve from https://digikogu.taltech.ee/en/Download/21e64f91-1694-4a9b-bb66-c2be28d3231a.

[10] Hupu, Retrieved from https://bbs.hupu.com/39307300.html.

[11] Basketball Reference, Retrieved from https://www.basketball-reference.com/.