|  |
| --- |
| **National Basketball Association salary prediction: a data-driven Linear Regression analysis** |

Jiahe Zhang

freddie\_1534@163.com

***Abstract.***

National Basketball Association (also known as NBA which will be used as an abbreviation throughout this report) as one of the few most successful professional sports leagues in the world, it is 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. This report 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 report reveals which specific variables are useful for predicting players’ salary using Multiple Linear Regression

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. While, to make the competition fair, there is a limit on the amount of money that a team can spend on players’ salaries, which is so called ‘salary cap’, if such limit was reached, luxury tax is required to pay towards the association. Therefore, making the budget manageable and spending the money efficiently is crucial. This is where a prediction algorithm comes 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.

The structure of this report is as follows: Literature review, Methodology, Conclusion, References

1. **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 [1]. There have already been some great papers researching on the factors influencing NBA players’ salary. A paper in 2018 claimed that the determining factors of NBA players' pay are experience, points, rebounds, assists and fouls, whereas 3-point shots made and Hollinger's player efficiency rating (PER) are insignificant [2]. Also, [3] aims to explore the best player selection strategies by studying the player statistics, team performance, and the salary cap. It explains basic NBA statistics concepts and how the players' efficiency is measured and their relationship with the team performance. However, there are only 450 players in NBA in 2021, it is relatively low (compared with 1,696 players in The National Football League, 780 players in Major League Baseball), it is 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 confirmed in [4], 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.

1. **Methodology**

**3.1 Data pre-processing**

It was researched in [6], a NBA player performance scoring system created by the author. 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. 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 at prime(12)Game Score in playoffs at prime (13)Win Share in regular seasons at prime (14)Win Share in playoffs at prime; 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 [5] and [6] need filtration. As the players come from very different periods, the model would not be accurate considering inflation, new salary clauses and changes on salary caps. Hence, this report only studies on the top 50 players from the year of 2000 to the year of 2022. After filtration and integration, data in Figure 1 was used throughout this report:

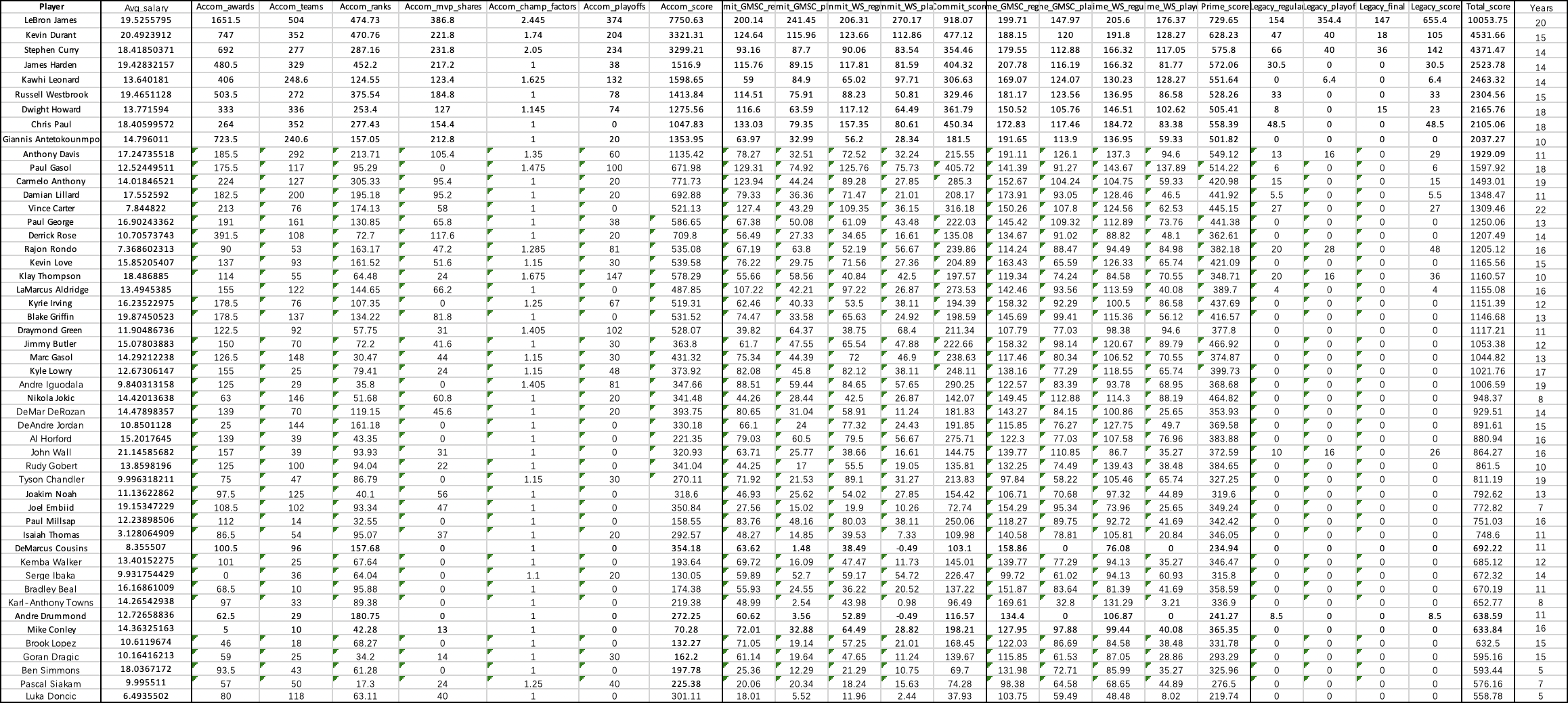


Figure 1: Top 50 players’ data

Figure 2 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’.

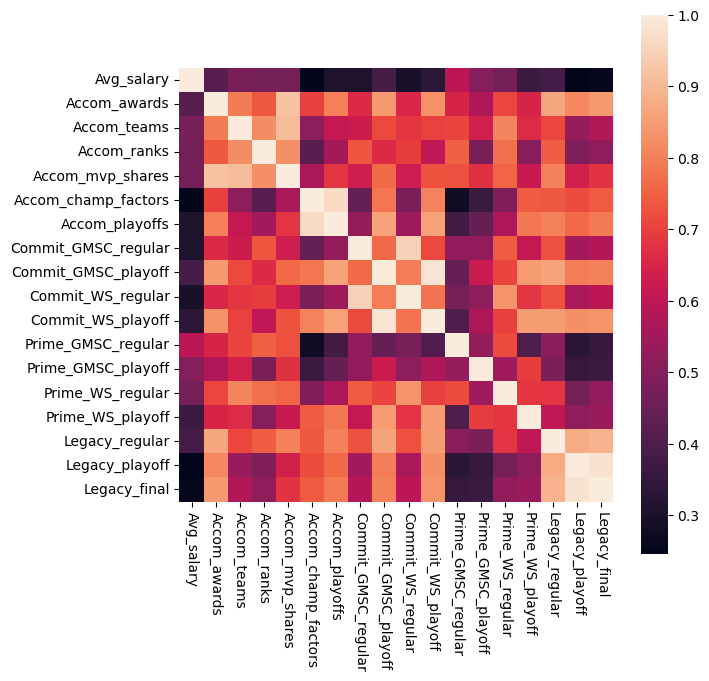


Figure 2: Correlation Matrix

**3.2 Prediction**

There were 2 regression models used in this report: Simple Linear Regression and Multiple Linear Regression

* + 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 report. The first model starts with Average Salary versus Total Score:

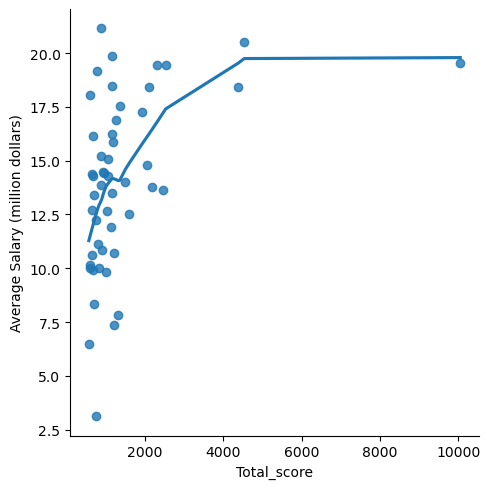


Figure 3: Simple Linear Regression results

Figure 3 contains the scattered data and the curve describing it, which indicates that there is not a strong linear relationship between Average Salary and Total Score.

Size of train set = **40** (80% of dataset)

Size of test set = **10** (20% of dataset)

Mean Squared Error 11.97

Size of train set = **45** (90% of dataset)

Size of test set = **5** (10% of dataset)

Mean Squared Error **19.62**

Figure 4: two results from Simple Linear Regression

Figure 4, 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 [5], 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.

* + 1. **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 shows the content each predictor helps explain the dependent variable. Figure 5 was attained after transforming and fitting Multiple Linear Regression of all 17 variables against Average Salary:



Figure 5: Multiple Linear Regression predictors information

In Figure 5, coef(β) refers to the amount Average Salary changes for one unit increase in predictor Xi; std err refers to Standard Error; t-statistic refers to indicates the linearity of each predictor with Average Salary; p-value indicates the correlation of each predictor with Average Salary. With Figure 5 and the formula, Average Salary = β0 + β1\* ‘Accom\_awards’ + β2 \* ‘Accom\_teams’ + β3\* …. +β4 \* ‘Legacy\_final’ + 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, ‘Commit\_WS\_regular’ has high p-value of 0.93, which indicates that it does not have a linear relationship with the dependent variable. Variables like this need to be dropped one by one and fitting the model again so that the model can be optimised, such method is so called Backward Selection. Average p-value was 0.3238 with all 17 variables used. After 9 times of Backward Selection, the variables left are (5)championship factors, (6)playoffs, (7)Game Score in regular seasons, (8)Game Score in playoffs, (10)Win Share in playoffs, (11)Game Score in regular seasons at prime, (14)Win Share in playoffs, (17)all-time leading records in finals. The lowest average p-value was obtained of value approximately 0.0747, which decreased 76.93% from 0.3238. While the Mean Squared Error was 7.9, which declined 34% from 11.97 and 59.7% from 19.62.

1. **Conclusion**

The performance of Multiple Linear Regression with Backward Selection was much better than that of Simple Linear Regression for the reason that Backward Selection eliminated the variables that were valuable for predicting salaries. Multiple Linear Regression had a result of average p-value 0.0747 and Mean Squared Error 7.9. And the most useful 8 variables, among the original 17 variables, predicting salary were: 1. championship factors, 2. playoffs, 3. Game Score in regular seasons, 4. Game Score in playoffs, 5. Win Share in playoffs, 6. Game Score in regular seasons at prime, 7. Win Share in playoffs, 8. all-time leading records in finals.

However, there are some other factors affecting the accuracy 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. Voluntary Pay cut: There were few some players who volunteer to have pay cut on themselves so that the team can sign new players. 3. Subjective factor: Decision makers were not always right. It has been highly common that players signing contracts with astronomical figure but have ordinary performance on court.

**References:**

1. Alyssa Schroer, ‘How Sports Analytics Are Used Today, by Teams and Fans’, 2018 (updated with additional reporting by Brian Nordli in 2021.)
2. Kevin Sigler, William Compton "NBA Players’ Pay and Performance: What Counts?", United States Sports Academy, 2018
3. R. Nagarajan and L. Li, "Optimizing NBA Player Selection Strategies Based on Salary and Statistics Analysis," 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, USA, 2017, pp. 1076-1083, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.175.
4. Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, Held August 24-26, 2021. Volume 2. Switzerland: Springer International Publishing, 2021.
5. <https://www.basketball-reference.com/players/>
6. 博闻爱打球, ‘NBA史上球员排行’, <https://bbs.hupu.com/39307300.html>, 2020