What really determins the salaries for NBA players?

Yiliu Cao

2022-12-17

Introduction

The NBA [1] is a top sporting event in North America, and the players of the NBA are also highly regarded. One of the reasons is that players in the NBA always have a high salary, and some top players have incredibly high salaries.

It is interesting to know why some players deserve high salaries but others do not. In this analysis, I will predict the NBA player's salaries by mainly focusing on the players' performance on the field. Therefore, my research question will be:**What determines an NBA player's salary, and how can we measure NBA players' salary based on their age and their performance on the field (like Assists and Points per game, etc.)?**

There were some academic papers investigating a similar question. Lyrons JR et al. (2015) [2] and Sigler et al. (2000) [3] argue that the points per game, field goal rate, and rebounds per game may significantly influence a player's salary. For my analysis, I agree that the three attributes they argued would influence the salary, and I will also include them in my model. However, they only focus on the player's in-game performance, but I believe the player's age is also a potential factor. Thus I will consider the player's ages when predicting their salaries.

Methods

In the section, I will introduce the methods we need to predict NBA players' salaries using linear regression.

The first thing is to divide the original data set into a training and a test data set, each containing 50% of the original data set. We will perform most analyses in the training data set. After that, we need to select the variables we will use and clean our data (e.g., remove NA observations).

Once we select all the variables, we can implement the Exploratory Data Analysis (EDA). Then we can build the full model in training data using all variables. For the full model, we first plot the response against the fitted values to check the additional condition 1 and scatterplot of all the predictors for additional condition 2. If any conditions do not hold, the patterns we see later when checking the assumptions of linear regression may not tell us what is wrong. If both additional conditions hold, we then check the assumptions of linearity, uncorrelated error, and constant variance by plotting the residuals versus fitted values and each predictor and checking the normality assumption by a Normal QQ plot.

If there are any violations of assumptions, we can use Box-Cox transformation to find appropriate transformation on predictors and response and implement what Box-Cox suggest. After we finish the transformation, we still need to check the two additional conditions and four assumptions and perform transformations until there are no severe violations of assumptions.

If there are no violations of assumptions, we may consider reducing the model. If you consider reducing the model. We will start by removing the predictors showing multicollinearity one by one and repeat this process until we find all reserved predictors are not correlated with each other. After that, we can look for candidate models. We can simply take the model we just have, which contains no multicollinearity, or we can also use the T-test to only choose the significant predictors from the transformed full model. However, the key things are that the candidate models should not have multicollinearity, and it is appropriate to remove all other predictors (by partial F test). Again, we still need to check the additional conditions and assumptions for each candidate model.

If you do not reduce the model or already have some Candidate models, we now need to perform diagnostics, including checking the multicollinearity and finding all problematic points (Leverage, Outlier, and Influential points). If the model we have now still indicates multicollinearity, we may need to change the model and recheck the additional condition and assumptions. Besides, if there are any problematic points, and it is necessary and ethical to remove them, we need to do the entire process again since we change our data set. Also, we can calculate the value of R squared, adjusted R squared, AIC, and BIC to access the goodness of the candidate models.

The last part is the Validation process. We need to apply the same transformations and fit the same models on the test data set as we did in the training data set. We then need to compare the properties of each model in each data set, including but not limited to the significance of coefficients, assumptions, multicollinearity, problematic points, etc. We will say our model is validated if our preferred models look similar to how they performed in the train data set. Otherwise, the model is not validated.

Results

For my data set, there are in total of 1408 observations, and thus there will be 704 observations for the training and the test data sets. However, there are missing values in each data set, and I have to remove them to do any further analysis. Therefore, the actual number of observations in the two data sets is 663 and 640, respectively. Besides, I only kept those 14 out of the above 40 variables. You can find more details in **Appendix 1)**.

Below are the numerical and graphical summaries of each data set.

Table 1: The numberical summaries for training data sets

Variables	Minimum	Median	Maximum	Mean	Variance	IQR
Age	19.0	26.00	41.00	26.27	1.877000e+01	6.00
Minutes Played(MP)	2.3	19.20	37.20	20.13	7.625000e+01	13.90
Field Goals per Game(FG)	0.1	2.70	10.80	3.23	4.610000e+00	2.80
Effective Field Goal Percentage(eFG.)	0.1	0.51	0.76	0.51	1.000000e-02	0.08
Free Throws(FT)	0.0	1.00	9.70	1.43	1.930000e+00	1.40
Free Throws Percentage(FT.)	0.0	0.76	1.00	0.74	2.000000e-02	0.14
Total Rebound per Game(TRB)	0.2	3.20	15.20	3.73	6.200000e+00	2.80
Assists per Game(AST)	0.0	1.30	10.70	1.91	3.030000e+00	1.75
Steals per Game(STL)	0.0	0.50	2.20	0.64	1.600000e-01	0.50
Blocks per Game(BLK)	0.0	0.30	2.70	0.42	1.800000e-01	0.35
Turnovers per Game(TOV)	0.0	0.90	5.40	1.14	6.300000e-01	0.90
Personal Fouls per Game(PF)	0.0	1.70	3.80	1.75	5.200000e-01	1.10
Points per Game(PTS)	0.4	7.10	36.10	8.71	3.587000e+01	7.50
Salary	77250.0	3627842.00	34682550.00	6766091.09	5.199913e+13	8480767.50

Table 2: The numberical summaries for test data sets

Variables	Minimum	Median	Maximum	Mean	Variance	IQR
Age	19.0	25.00	42.00	26.04	1.822000e+01	6.00
Minutes Played(MP)	2.9	21.25	37.80	21.31	7.632000e+01	13.83
Field Goals per Game(FG)	0.0	3.00	10.50	3.51	4.790000e+00	3.00
Effective Field Goal Percentage(eFG.)	0.0	0.51	0.88	0.51	1.000000e-02	0.07
Free Throws(FT)	0.0	1.10	9.20	1.53	1.820000e+00	1.32

Variables	Minimum	Median	Maximum	Mean	Variance	IQR
Free Throws Percentage(FT.)	0.0	0.77	1.00	0.74	2.000000e-02	0.15
Total Rebound per Game(TRB)	0.0	3.35	16.00	3.90	6.190000e+00	2.90
Assists per Game(AST)	0.0	1.50	11.20	2.06	3.410000e+00	1.90
Steals per Game(STL)	0.0	0.60	2.40	0.68	1.700000e-01	0.50
Blocks per Game(BLK)	0.0	0.30	2.40	0.42	1.600000e-01	0.40
Turnovers per Game(TOV)	0.0	1.00	5.70	1.19	6.600000e-01	1.00
Personal Fouls per Game(PF)	0.0	1.80	3.90	1.80	5.000000e-01	0.80
Points per Game(PTS)	0.5	7.80	29.10	9.47	3.716000e+01	8.03
Salary	56845.0	3522560.00	37457154.00	7134345.04	5.953722e+13	8985633.00

Table 1 and 2 show the numerical summaries of both data sets. We can see that the two data sets have a similar performance on almost every variable. However, the minimum total rebound per game is 0.2 in the training data set but 0 in the test data set. Also, the IQR of the Salary in the test data set is about 1.5 million higher than in the training data set.

Figure 1: The graphical summaries of Salary in training and test data set

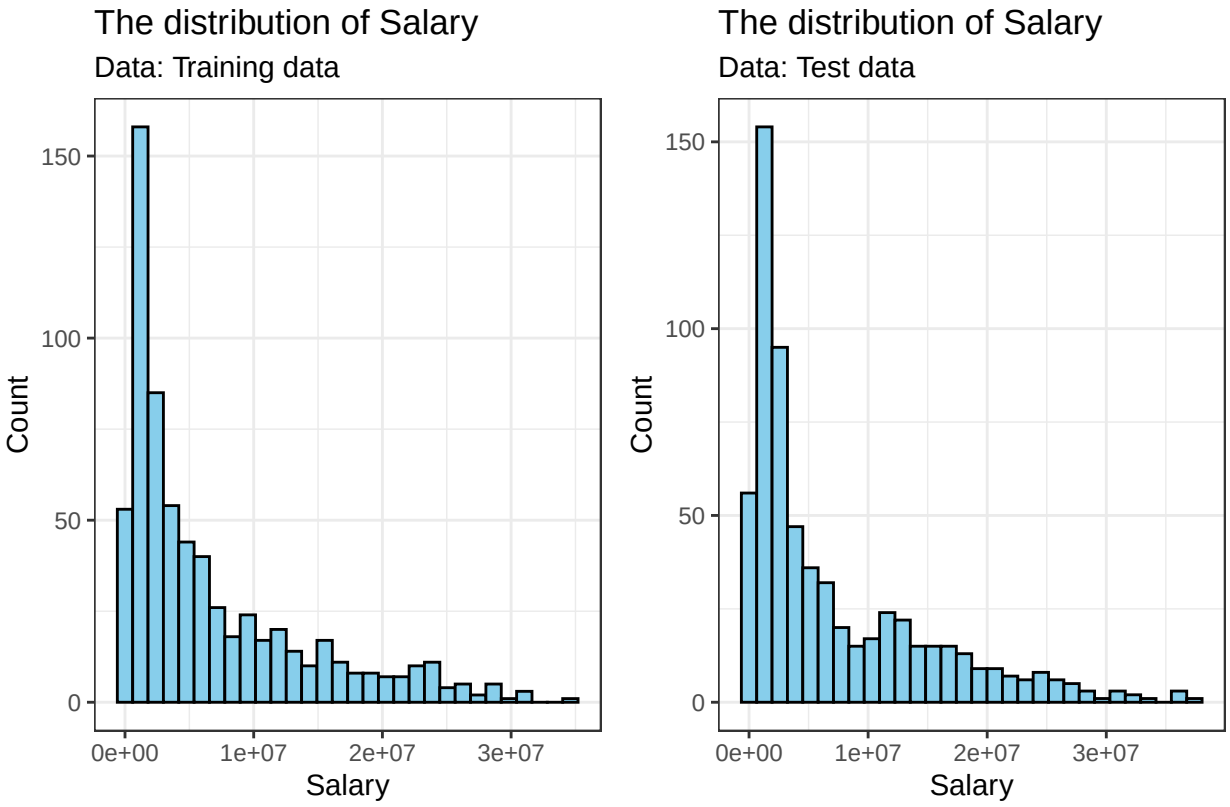

Figure 1 shows the distribution of salary in both data sets. We can observe that they are both right-skewed, and we can also observe the salary has a higher IQR in the test data set.

We will start with the full model in the training data set, in which all variables in Table 1 are predictors with "Salary" as the response. It satisfies both additional conditions but violates the assumption of constant variance and normality. Therefore we need to do the transformations on both the response and the predictors by Box-Cox (**Appendix 2)**). For the transformed model, we found that assumption of constant variance is still violated. Now should try to reduce our model by removing the predictors showing multicollinearity one by one, and eventually, there are two candidate models. Again, we still need to check the additional assumptions for the two candidate models and assumptions. However, we can see that the two models still have violations of constant variance.

Next is to perform a diagnostic on these two models.

Table 3: The summary of goodness of the full model and the two Candidate models

Model	R^2	Adjusted R^2	AIC	BIC	Largest VIF
Full model	0.5819082	0.5735334	1.0840047×10^4	1.0907498×10^4	160.6644498
Candidate model 1	0.5780104	0.5721943	1.0838199×10^4	1.0887664×10^4	4.6247901
Candidate model 2	0.5764242	0.5712428	1.0838687×10^4	1.0883654×10^4	4.717484

Table 3 summarizes the goodness of the transformed full model and the two Candidate models. We can see that the Candidate models have a lower value of AIC and BIC than the full model and a lower multicollinearity.

Up to now, we have built two models and performed the diagnostic. We should do the validation process now, and the results are shown below.

Table 4: Summary of characteristics of two candidate models in the training and test data sets (Response: $\sqrt{Salary}$). Coefficients are presented as estimate ± SE (* = significant t-test at $\alpha = 0.05$)

Characteristic	Model1 (Train)	Model1 (Test)	Model 2 (Train)	Model 2 (Test)
R^2	0.5780104	0.5708705	0.5764242	0.5597621
Adjusted R^2	0.5721943	0.56474	0.5712428	0.5541807
AIC	1.0838199×10^4	1.0541548×10^4	1.0838687×10^4	1.0555904×10^4
BIC	1.0887664×10^4	1.0590624×10^4	1.0883654×10^4	1.0600518×10^4

Characteristic	Model1 (Train)	Model1 (Test)	Model 2 (Train)	Model 2 (Test)
Largest VIF value	4.6247901	4.4032141	4.717484	4.4360816
# Leverage points	39	42	40	39
# Outliers	0	0	0	0
# Cook's D	0	0	0	0
# DFFITS	39	30	34	28
# DFBETAS	24	20	20	21
Violations	Constant Variance	Constant Variance & Normality	Constant Variance	Constant Variance & Normality
Intercept	3882.722 ± 394.816 (*)	3316.014 ± 421.195 (*)	3862.6 ± 393.817 (*)	3194.253 ± 427.212 (*)
$1/Age$	-8.1767919×10^4 ± 5544.208 (*)	-7.9889086×10^4 ± 6105.443 (*)	-8.2152322×10^4 ± 5543.106 (*)	-7.9756965×10^4 ± 6190.762 (*)
$\sqrt{FG}$	-	-	1069.957 ± 122.615 (*)	800.087 ± 129.183 (*)
eFG.	-1761.679 ± 506.575 (*)	-1471.477 ± 555.036(*)	-1879.137 ± 508.372 (*)	-1423.904 ± 565.656(*)
$FT.^2$	-555.097 ± 222.252 (*)	-83.279 ± 250.262	-375.463 ± 215.481	105.531 ± 244.55
$\sqrt{TRB}$	607.835 ± 110.524 (*)	484.528 ± 110.759 (*)	556.095 ± 101.086 (*)	602.161 ± 101.676 (*)
$\sqrt{AST}$	73.686 ± 103.594	260.044 ± 103.983 (*)	82.836 ± 102.299	280.47 ± 103.344(*)
$\sqrt{STL}$	122.247 ± 210.164	134.289 ± 211.819	148.781 ± 209.212	231.886 ± 212.919
$\sqrt{BLK}$	-33.977 ± 178.727	364.608 ± 180.075 (*)	-	-
PF	-175.379 ± 76.456 (*)	-144.027 ± 79.939	-164.87 ± 74.972 (*)	-108.948 ± 79.037
$\sqrt{PTS}$	646.814 ± 72.895 (*)	553.233 ± 76.608 (*)	-	-

Table 4 shows the validation process for the two data sets and two models.

Comparing the general performance of the two models, model 2 seems to be better than model 1. Even though model 2 has a similar goodness to model 1, the percentage difference in estimates is much less than model 1. Therefore I will use model 2 to predict NBA players' Salaries.

Discussion

From previous parts, we have concluded the final model to predict the NBA players' Salaries is model 2. From the model 2, it tells us we can use a player's age, field goals per game, Effective Field Goal Percentage, Free Throw Percentage, Total Rebound per game, Assists per game, Steals per game, and Personal Fouls per Game to predict a player's salary. Besides, we expect a 164.87 decrease in the square root of salary for every unit increase in Personal Fouls per Game. This makes sense as those "aggressive" usually can not live long in the NBA and thus have a lower salary. However, remember we need transformations when we implement the model.

Now, we can answer our research question. The predictors I just stated will be highly influencing players' salaries. Using model 2 in Table 4, we can measure NBA players' salaries.

However, there are still some limitations in my analysis.

1. There is always a violation of the assumption of constant variance. Perhaps our data is small and has too many variations, but there are also a relatively large proportion of leverage points in our models. Hence, this may reduce the accuracy of our models, and the prediction of salaries may be less reliable.

2. The adjusted R squared is relatively small, and values of AIC and BIC are still relatively large, indicating that the goodness of our models still needs to be enhanced. The possible reason is that our models are not complicated enough. We may try to add more predictors.

3. The model transformation we perform by Box-Cox transformation may be too simple. This means that we used the rounded value of lambda from the Box-Cox transformation. This may result in some possible assumption violations as we did not follow precisely what Box-Cox suggests.

4. We can not validate the model. This means we can only use the information to understand the limitations better. The potential reasons are similar to the first one; our transformation may be too specific to training data and can not help with the test data.

In conclusion, we can use linear regression to predict NBA players' salaries. However, some limitations still exist and can be improved in future analyses.

Reference list

1. Wikimedia Foundation. (2022, December 2). National Basketball Association. Wikipedia. Retrieved December 20, 2022, from https://en.wikipedia.org/wiki/National_Basketball_Association

2. Lyons Jr, R., Jackson Jr, E. N., & Livingston, A. (2015). Determinants of NBA Player Salaries. The Sport Journal. https://doi.org/ 10.17682/sportjournal/2015.019

3. Sigler, K. J., & Sackley, W. H. (2000). NBA players: are they paid for performance? Managerial Finance, 26(7), 46–51. https:// doi.org/10.1108/03074350010766783

4. Analysis, P. (2020, January 10). Understanding basketball analytics: EFG% vs. FG%. PivotAnalysis. Retrieved December 20, 2022, from https://www.pivotanalysis.com/post/what-is-efg

Appendix

1). Justifications about the variables I removed from the original data set.
Below is the summary of variables which I removed initially.

Table 5: The summary of initally removed variables

Category 1	Category 2	Category 3
Player's name	The field goal attempt	Rk
Player's ID	The Filed goal rate	Fvot
The first position	3-Point Field Goals Per Game	FRank
The second position	3-Point Field Goal Attempts Per Game	Pvot
The player's team	FG% on 3-Pt FGA	PRank
The game played	2-Point Field Goals Per Game	Mvot
The game started	2-Point Field Goal Attempts Per Game	MRank
Daily views on wikipedia	FG% on 2-Pt FGA	Score
Seaon of NBA	Free Throw Attempts Per Game	
Player's conference	Offensive Rebounds Per Game	
Players's role	Defensive Rebounds Per Game	
If the player played in the all star game		

Table 5 shows the summary of variables that I removed initially.

For the variables in "Category 1": I remove them because they are useless for predicting a player's salary. For instance, we can not predict a player's salary according to his name.

For the variables in 'Category 2": The reason why I remove them is that they are redundant, and I reserve a similar variable. For instance, I keep the variable "Total Rebound per Game" so that it is unnecessary for me to still keep "Offensive Rebound per Game" and "Defensive Rebound per Game."

For the variables in "Category 3": The reason why I remove them is that the data source link does not provide what they mean. For instance, I do not know what "Rk" means or the difference between "FRank" and "PRank." You can refer to Table 1 or Table 2 for more details about the variables I kept. Besides, I choose to use an Effective Field Goal Rate instead of Field Goal Rate [4].

2). The results of Box-Cox transformation is:

Table 6: The summary of Box-Cox transformation in the training data set for the full model

Variables	Original form (Shorted Names)	After Transformation
Age	Age	$1/Age$
Minutes Played	MP	$\sqrt{MP}$
Field Goals per Game	FG	$\sqrt{FG}$
Effective Field Goal Percentage	eFG.	eFG.
Free Throws	FT	$\sqrt{FT}$
Free Throws Percentage	FT.	$FT.^2$
Total Rebound per Game	TRB	$\sqrt{TRB}$
Assists per Game	AST	$\sqrt{AST}$
Steals per Game	STL	$\sqrt{STL}$
Blocks per Game	BLK	$\sqrt{BLK}$
Turnovers per Game	TOV	$\sqrt{TOV}$
Personal Fouls per Game	PF	PF
Points per Game	PTS	$\sqrt{PTS}$
Salary	Salary	$\sqrt{Salary}$