Analyze NBA Players' Salaries Based on Multiple Regression

BIA-652 Final Project

Professor: Feng Mai

Group members: Chih-Po Wang, Chun-Wei Hung, Xinze Yu, Yun-Hao Tung

Section 1: Introduction & Problem Overview and Motivation

In statistics, multiple regression (also known as multiple linear regression) is a linear approach for modelling the relationship between a scalar response and many explanatory variables (also known as dependent and independent variables). There are two main advantages to analyzing data using a multiple regression model. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The second advantage is the ability to identify outliers, or anomalies. We are going to use multiple regression to analyze the technical factors that are influencing NBA players' salary.

NBA, one of the four major sports leagues in North America, is the highest level in the basketball field and the most successful internationally promoted one among four kinds of sports. As more and more streaming platforms pop up, fans from all over the world can enjoy watching live broadcasts of NBA games through TV, the Internet, and mobile devices. The change also reflects in the number of subscriptions to NBA League Pass, the official streaming platform of the NBA. Since 2017, the number of subscribers per year has double-digit growth annually. Besides, the global influence of the NBA also makes it a stage that all basketball players chase.

The salary cap of the NBA, which means the highest total salaries of a team, has risen from 70 million U.S. dollars in 2015 to 112 million U.S. dollars in 2021. And this change

is reflected in players' income as well. Many players have signed record-breaking contracts in the past six years.

Different from the past style, the current NBA team style has a more favorable offensive side. Three-point shooting and high-paced game rhythm have become mainstream. Even free throws have become primary scoring methods for players nowadays. Therefore, compared to the past when teams valued the ability to consolidate rebounds and two-point shots in the penalty area, defenders and Small Forwards with long-distance shooting capabilities and rapid advancement have become the cornerstones of each team. In the 2021 player salary rankings, only 4 of the top 20 highest-salary players serve as Power Forwards or Centers, and the rest are all Small Forwards and Guards.

Such an extreme distribution makes us wonder what abilities a player has will arouse the team's willingness to invest in high salaries.

Section 2: Problem Objective

The analysis of the factors that affect NBA players' salaries may help teams choose the right players more clearly and provide valuable insights for players seeking to increase their salaries.

We can also clearly understand the following issues:

1. Clarify the factors that affect NBA players' salaries.
2. Analyze what are the main factors affecting salaries.

Section 3: Dataset

Dependent variable and independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{15} X_{15} + \varepsilon$$

The values β_i represent parameters to be estimated, and ε is the independent identically distributed normal error.

The dependent variable:

CSS (Y): Player's current season salary

The independent variable:

AGE (X_1): The NBA players' ages when in each season

GS/G (X_2): The percentage of the NBA players are starting line-up in their current season

MP/G (X_3): The NBA players' minutes played time per game in current season

FG% (X_4): Player's field goal percentage

3P% (X_5): Player's 3-points field goal percentage

2P% (X_6): Player's 2-points field goal percentage

eFG% (X_7): Player's effective field goal percentage

FT% (X_8): Player's free throw percentage

TRB/G (X_9): Average rebounds that players have per games

AST/G (X_{10}): Average assists that players make per games

STL/G (X_{11}): Average steals that players make per games

BLK/G (X_{12}): Average blocks that players make per games

TOV/G (X_{13}): Average turnover that players make per games

PF/G (X_{14}): Average personal fouls that players make per games

PTS/G (X_{15}): Average points that players scored per games

Section 3.1: Preliminary Exploratory Data Analysis

Figure 1 shows the missing values percentage of our raw data, to avoid missing values, we use the mean value of each column to fill in the missing values.

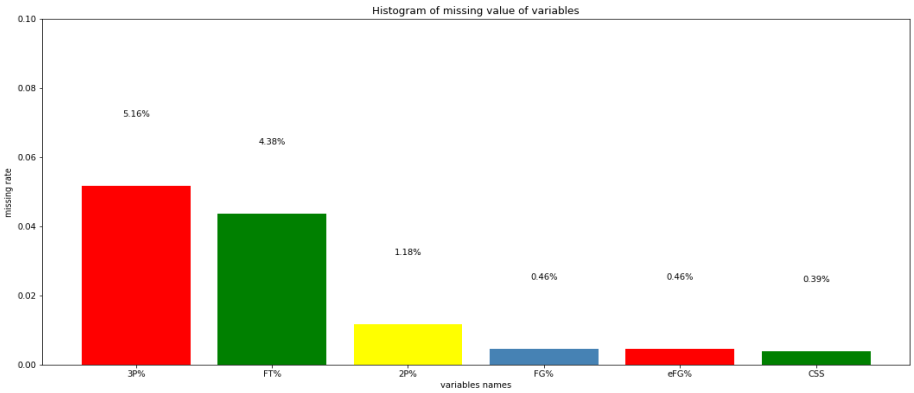

Figure 1: Histogram of missing value of variables

In order to test the accuracy of the model we built, we divided the data into 80% training data and 20% test data. After the model is seen from the train data, the test data is used to bring in the predicted salary (Y_p) compare with actual salary (Y_a).

Figure 2 shows the description of our preprocessed data structure (full-model).

	count	mean	std	min	25%	50%	75%	max
CSS	1224	67.1802568	86.9319671	0.11898	13.78242	27.85	87.5891075	700
Age	1224	25.6764706	4.12616666	19	22	25	28	42
GS/G	1224	0.36601225	0.39316387	0	0.016	0.173	0.7715	1
MP/G	1224	19.5206087	9.0972428	0.667	12.2875	19.2145	27.27475	37.563
FGP	1224	0.44743498	0.10333876	0	0.403	0.441	0.493	1
3PP	1224	0.31415063	0.12222044	0	0.286	0.335	0.378	1
2PP	1224	0.50898265	0.11438127	0	0.462	0.51	0.56525	1
eFGP	1224	0.51186923	0.10370893	0	0.479	0.52	0.55925	1.5
FTP	1224	0.73992104	0.14064099	0	0.682	0.757	0.826	1
TRB/G	1224	3.62329085	2.44653969	0	1.8965	3.1855	4.7325	15.595
AST/G	1224	1.90121487	1.76778752	0	0.73575	1.3305	2.4205	10.795
STL/G	1224	0.61146487	0.39733399	0	0.325	0.556	0.85125	2.208
BLK/G	1224	0.4009085	0.4056897	0	0.136	0.283	0.532	3.383
TOV/G	1224	1.07396895	0.80640687	0	0.52125	0.868	1.4	5
PF/G	1224	1.71282843	0.78787105	0	1.1345	1.7195	2.24525	4.875
PTS/G	1224	8.64500735	6.24856143	0	4.09325	7.0615	11.5535	36.128

Figure 2: Data description

Also, Figure 3 shows the distribution of Y.

For example, As can be seen from the figure below, the current season salary follows a right skewed distribution.

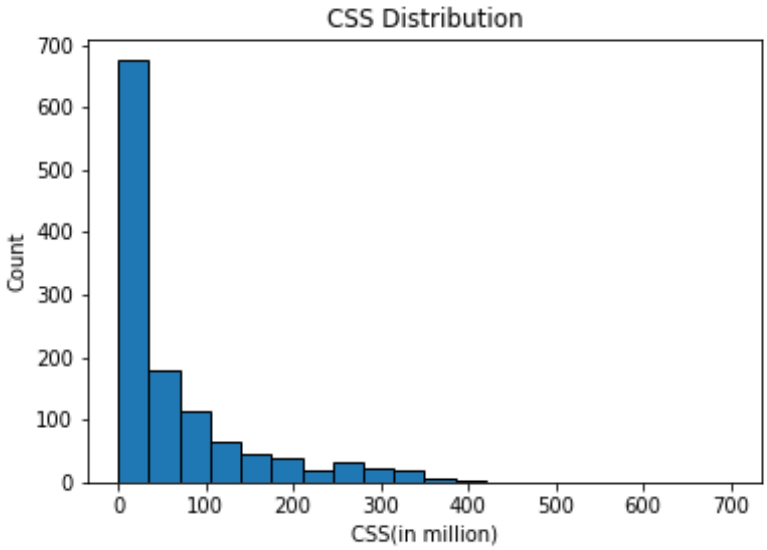

Figure 3: CSS distribution

It can be seen from the Figure 4 below that among the factors that affect players' salary, the PTS/G and PF/G are highly correlated; TRB and FT% is moderately correlated; STL and Age is negatively correlated.

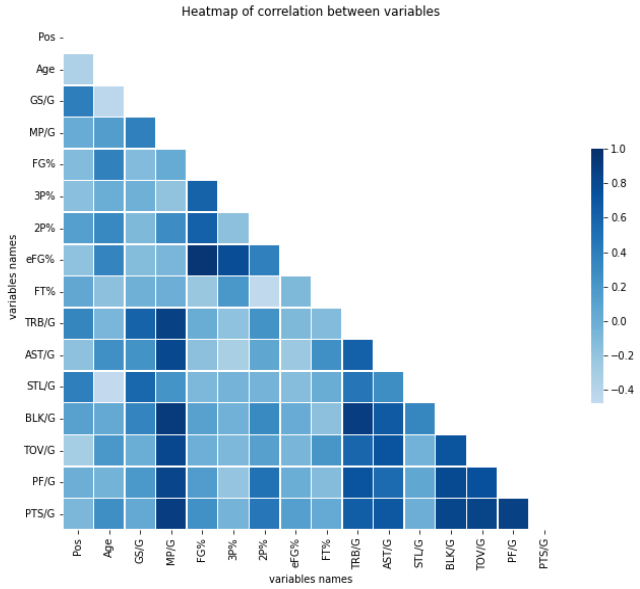

Figure 4: Heatmap of correlation between variables

Section 4.1: Fit with Full Model

Since the number of Y (current season salary) is too large, we take million as the unit for Y in regression.

$$E(\widehat{Y}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \varepsilon_i, \quad i = 1, 2, 3, \dots, 1224$$

To test the above basic hypothesis, we include all variables that may have an impact as explanatory variables.

OLS Regression Results						
Dep. Variable:	CSS	R-squared:	0.594			
Model:	OLS	Adj. R-square	0.589			
Method:	Least Square	F-statistic:	117.6			
Date:	Sun	12-Dec-21	Prob (F-stati:	1.52E-223		
Time:	5:39:52	Log-Likelihoc	-6650.6			
No. Observat	1224	AIC:	1.33E+04			
Df Residuals	1208	BIC:	1.34E+04			
Df Model:	15					
Covariance T	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-139.5721	16.233	-8.598	0	-171.421	-107.723
Age	6.5141	0.41	15.895	0	5.71	7.318
GS/G	51.8947	7.419	6.995	0	37.34	66.449
MP/G	-2.8884	0.575	-5.021	0	-4.017	-1.76
FGP	-60.9714	54.682	-1.115	0.265	-168.253	46.31
3PP	-16.1308	18.034	-0.894	0.371	-51.511	19.25
2PP	-2.9855	21.877	-0.136	0.891	-45.907	39.936
eFGP	36.2088	54.215	0.668	0.504	-70.158	142.576
FTP	5.9972	13.05	0.46	0.646	-19.607	31.601
TRB/G	7.8283	1.271	6.16	0	5.335	10.321
AST/G	8.493	1.941	4.377	0	4.686	12.3
STL/G	13.051	6.402	2.038	0.042	0.49	25.612
BLK/G	8.4156	5.892	1.428	0.153	-3.143	19.975
TOV/G	-0.7355	5.028	-0.146	0.884	-10.599	9.128
PF/G	-15.1974	3.612	-4.207	0	-22.284	-8.11
PTS/G	6.6104	0.673	9.821	0	5.29	7.931
Omnibus:	512.578	Durbin-Wats	1.883			
Prob(Omnibu	0	Jarque-Bera	8669.544			
Skew:	1.488	Prob(JB):	0			
Kurtosis:	15.694	Cond. No.	1.65E+03			
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[2] The cond	1.65e+03. This might indicate that there are					
strong multicollinearity or other numerical problems.						

Figure 5: Full model regression results

Obtained from the figure 5 above:

$$b_0 = -139.6, b_1 = 6.5, b_2 = 51.9, b_3 = -2.9, b_4 = -61.0, b_5 = -16.1,$$
$$b_6 = -3.0, b_7 = 36.2, b_8 = 6.0, b_9 = 7.8, b_{10} = 8.5, b_{11} = 13.1, b_{12} = 8.5,$$
$$b_{13} = -0.7, b_{14} = -15.2, b_{15} = 6.6$$

$$\hat{Y} = -139.6 + 6.5x_1 + 51.9x_2 - 2.9x_3 - 61.0x_4 - 16.1x_5 - 3.0x_6 + 36.2x_7 + 6.0x_8 + 7.8x_9 + 8.5x_{10} + 13.1x_{11} + 8.5x_{12} - 0.7x_{13} - 15.2x_{14} + 6.6x_{15}$$

Section 4.2: Parameter test

After getting the β values, we want to judge whether there is a linear relationship between each explanatory variable X_1~X_{15}.

For example, to determine whether there is a linear relationship between **Age (X_1)** and salary (Y), we assume hypothesis testing:

$$\alpha = 0.1$$
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

According to the results, since $P - value < \alpha = 0.1$, therefore we reject H_0, which means that we have enough evidence to show that $\beta_1 \neq 0$.

That means Age (X_1) has a linear relationship with salary (Y).

Under the circumstance that variables other than Age are fixed, when age increases by 1 year, the current season salary increases by $651,400. For the rest of the variables, we apply the same methods to them. The results show that the following variables are linearly related to y(current season salary):

Age (X_1); Percentage of starting line-up (X_2); Minutes played time per game (X_3)； Average rebounds per games (X_9)； Average assists per games (X_{10})；

Average steals per games (X_{11})； Average personal fouls per games (X_{14})； Average points per games (X_{15}).

Also, some variables no linear relationship with y(current season salary):

Field goal percentage (X_4); 3-points field goal percentage (X_5); 2-points field goal percentage (X_6); effective field goal percentage (X_7); free throw percentage (X_8); Average blocks per games (X_{12}); Average turnover per games (X_{13}).

Section 4.3: Model Interpretation Ability

Within the full-model and alpha = 0.1, we find that some variables are linearly related, but some are not. Via the GOF (goodness of fit) test, we know that the entire variables can explain y (current season salary). Also, we find the R^2 of the full model is 0.594.

Section 5: Model Selection

We hope to get the most accurate and concise model. So we first use VIF to drop some highly correlated features.

Section 5.1: Collinearity Analysis based on VIF

	feature	VIF
0	const	103.727826
1	Age	1.11600797
2	GS/G	3.42783624
3	MP/G	10.9059423
4	FGP	13.4284225
5	3PP	2.12999058
6	2PP	2.48333937
7	eFGP	13.623846
8	FTP	1.28201886
9	TRB/G	3.84172126
10	AST/G	4.81971062
11	STL/G	2.65616837
12	BLK/G	2.19816694
13	TOV/G	6.69938564
14	PF/G	3.21540792
15	PTS/G	7.1887533

Figure 6: VIF of full model

As can be seen from the above figure 6, the variance expansion coefficient (VIF) of the minutes played time per game (X_3), Field goal percentage (X_4), and effective field goal percentage (X_7), is greater than 10, indicating these three variables and other variables. There is a problem of collinearity, so we consider removing these three variables to solve the problem of collinearity.

Section 5.2: Forward Selection Method

When selecting variables, we use the forward selection method and other selection methods (adjusted R^2) to select the most suitable combination of variables and the following is the selection method we used.

The forward selection method is to add forecast variables one by one. If this forecast variable has better explanatory power, add it to the model. If other remaining forecast variables have better explanatory power than variables that have been added to the model in the previous step, then add the model, until the last remaining predictive variables do not have sufficient partial explanatory ability.

	feature_idx	cv_scores	cv_scores	feature_names	ci_bou	std_de	std_er	r-square	Adjusted r-square
1	(11,)	3965.5379590	-3337.81651357 -	('PTS/G',)	1089	847	424	0.430972514	0.430506861
2	(0, 11)	3189.0905386	-2631.17815097 -	('Age', 'PTS/G')	1065	828	414	0.530499143	0.5297301
3	(0, 1, 11)	3079.2512633	-2725.03254728 -	('Age', 'GS/G', 'PTS	984	765	383	0.548485432	0.547375151
4	(0, 1, 10, 11)	3033.8478084	-2637.25085956 -	('Age', 'GS/G', 'MP	1020	794	397	0.555214063	0.553754552
5	(0, 1, 5, 10, 11)	2977.8788122	-2655.83050235 -	('Age', 'GS/G', 'MP	971	756	378	0.571029476	0.569268513
6	(0, 1, 5, 6, 10, 11)	2971.0074420	-2426.13955788 -	('Age', 'GS/G', 'MP	1068	831	415	0.582072373	0.580011924
7	(0, 1, 2, 5, 6, 10, 11)	2891.9130378	-2473.45899012 -	('Age', 'GS/G', 'MP	1039	808	404	0.583280428	0.580881549
8	(0, 1, 2, 3, 5, 6, 10, 11)	2901.3182357	-2462.35767196 -	('Age', 'GS/G', 'MP	1039	808	404	0.583571083	0.580536596
9	(0, 1, 2, 3, 5, 6, 8, 10, 11)	2899.7071649	-2458.0167667 -	('Age', 'GS/G', 'MP	1030	801	401	0.58475135	0.580483883
10	(0, 1, 2, 3, 4, 5, 6, 8, 10, 11)	2899.7071649	-2458.0167667 -	('const', 'Age', 'GS,	1030	801	401	0.584751434	0.581328031
11	(0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11)	2900.4026176	-2458.5808775 -	('const', 'Age', 'GS,	1030	802	401	0.58479016	0.580982677
12	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	2900.8551363	-2461.75049079 -	('const', 'Age', 'GS,	1033	803	402	0.584957531	0.580675776

Figure 7: Forward Selection Method

The adjusted coefficient considers the index of the degree of freedom.

1. As long as the new predictive variable has sufficient explanatory power, the SSE value will decrease. And the degree of SSE decline can be compensated by the loss of the degree of freedom, then the adjusted coefficient will increase.

2. If the explanatory power of the new forecast variable is not enough, the declination of the degree of SSE cannot be compensated by the lost degree of freedom, and that makes the adjusted coefficient drop.

Therefore, we can determine whether the predicted variable should be added to the model based on whether the adjusted determination coefficient has increased significantly or not.

Section 5.3: Model confirmation

After using the forward selection method and using the adjusted R^2, variables Age, GS/G, 3PP, 2PP, FTP, TRB/G, AST/G, BLK/G, PF/G, PTS/G are selected.

Thus, the final model will be:

$$E(\hat{Y}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{12} x_{i12} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \varepsilon_i, \qquad i = 1, 2, 3, \ldots, 1224$$

Section 5.4: Fitness test

OLS Regression Results						
Dep. Variable:	CSS	R-squared:	0.585			
Model:	OLS	Adj. R-squared:	0.581			
Method:	Least Square	F-statistic:	170.8			
Date:	Sun	12-Dec-21	Prob (F-statistic):	2.15E-223		
Time:	5:41:25	Log-Likelihood:	-6663.7			
No. Observations:	1224	AIC:	1.34E+04			
Df Residuals:	1213	BIC:	1.34E+04			
Df Model:	10					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-142.9838	15.636	-9.145	0	-173.66	-112.308
Age	6.3381	0.401	15.815	0	5.552	7.124
GS/G	37.5851	6.732	5.583	0	24.377	50.794
3PP	-24.356	13.433	-1.813	0.07	-50.711	1.999
2PP	-17.0765	15.043	-1.135	0.257	-46.59	12.437
FTP	-0.2029	12.959	-0.016	0.988	-25.627	25.221
TRB/G	6.7311	1.207	5.577	0	4.363	9.099
AST/G	7.8466	1.39	5.647	0	5.12	10.573
BLK/G	10.7819	5.84	1.846	0.065	-0.676	22.24
PF/G	-22.5817	3.143	-7.186	0	-28.747	-16.416
PTS/G	5.0586	0.51	9.924	0	4.059	6.059
Omnibus:	504.897	Durbin-Watson:	1.878			
Prob(Omnibus):	0	Jarque-Bera (JB):	8177.037			
Skew:	1.471	Prob(JB):	0			
Kurtosis:	15.316	Cond. No.	348			
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						

Figure 8: Selected model regression result

It can be seen from the above figure 23 that the minutes played time per game (X_3), Field goal percentage (X_4), effective field goal percentage (X_7), Average steals per game (X_{11}), Average turnover per games (X_{13}) have been deleted.

Section 5.5: Parametric test for model after selection

After getting the β values, we want to judge whether there is a linear relationship between each explanatory variable X_1~X_{15}.

For example, to determine whether there is a linear relationship between **Age (X_1)** and salary (Y), we assume hypothesis testing:

$$\alpha = 0.1$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

According to the results, since $P - value < \alpha = 0.1$, therefore we reject H_0, which means that we have enough evidence to show that $\beta_1 \neq 0$.

That means Age (X_1) has a linear relationship with salary (Y).

Under the circumstance that variables other than Age are fixed, when age increases by 1 year, the current season salary increases by $633,800.

For the rest of the variables, we apply the same methods to them.

The results show that the following variables are linearly related to y(current season salary):

Age (X_1); Percentage of starting line-up (X_2); 3-points field goal percentage (X_5); Average rebounds per game (X_9); Average assists per game (X_{10}); Average blocks per game (X_{12}); Average personal fouls per game (X_{14});

Average points per game (X_{15});

Also, some variables no linear relationship with y(current season salary):

2-points field goal percentage (X_6); Free throw percentage (X_8).

Section 5.6: Explanatory ability of model after selected

Under the five-variable model, R2=0.585, which means that the regression model can explain 58.5% of the model.

The difference with R2=0.594 under the original full model is only 0.009.

Section 6: Residual Analysis

The three main residual analysis assumption include test of normality, test of independence, and test of homogeneity of variance.

Durbin-Watson:	1.878

The Durbin Watson is used to detect whether the error term meets the residual analysis. Since Durbin Watson value = 1.878, we have enough evidence that the residuals meet the three tests.

Section 7: Predict the testing data

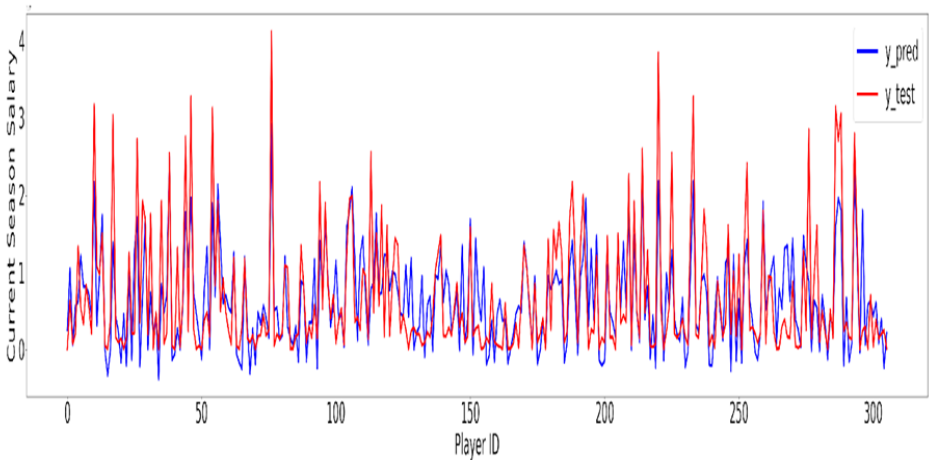

Figure 9: Line graph of testing data and data predicted from model

The red line in the graph above represent the real player salary, and the blue line represent the predicted salary obtained from variables through the model we select.
It can be seen from the graph that the distribution of the two lines is very similar. The salary of the players under the model prediction is a little bit lower than the real salary.

Section 8: Conclusion

Our report mainly discusses the impact of these 15 variables on player salaries. We delete unsuitable variables in the coefficient of variance expansion and forward selection to get the best combination of variables.
Through multiple linear regression, we get the following regression model

$$\hat{y} = -142.9838 + 6.3381x_1 + 37.5851x_2 - 24.356x_5 - 17.0765x_6 - 0.2029x_8 + 6.7311x_9 + 7.8466x_{10} + 10.7819x_{12} - 22.5817x_{14} + 5.0586x_{15}$$

And finally, this regression model is used for residual analysis, which meets the three major assumptions of residuals: normality, independence and homogeneity, so this regression model is the best linear regression model we can find.

Through regression analysis, we can understand that with the development trend of the NBA, in the past, people would think that players do not need to practice shooting well to get more salaries, but in fact it is just the opposite, and it should be when the player owns Better defense is easier to get a higher salary, so on the player's side, players should train more defensive skills so that they can make more money.

Reference

[1] NBA Player Salaries. *Hoopshype. Retrieved from* https://hoopshype.com/salaries/players/

[2] NBA Contracts. *spotrac. Retrieved from* https://hoopshype.com/salaries/players/

[3] Player Totals. *Basketball reference*. *Retrieved from* https://www.basketball-reference.com/