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

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 [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.

3. 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:

Player	Avg_salary	Accom_awards	Accom_teams	Accom_ranks	Accom_mvp_shares	Accom_champ_factors	Accom_playoffs	Accom_score
LeBron James	19.5255795	1651.5	504	474.73	386.8	2.445	374	7750.63
Kevin Durant	20.4923912	747	352	470.76	221.8	1.74	204	3321.31
Stephen Curry	18.41850371	692	277	287.16	231.8	2.05	234	3299.21
James Harden	19.42832157	480.5	329	452.2	217.2	1	38	1516.9
Kawhi Leonard	13.640181	406	248.6	124.55	123.4	1.625	132	1598.65
Russell Westbrook	19.4651128	503.5	272	375.54	184.8	1	78	1413.84
Dwight Howard	13.771594	333	336	253.4	127	1.145	74	1275.56
Chris Paul	18.40599572	264	352	277.43	154.4	1	0	1047.83
Giannis Antetokounmpo	14.796011	723.5	240.6	157.05	212.8	1	20	1353.95
Anthony Davis	17.24735518	185.5	292	213.71	105.4	1.35	60	1135.42
Paul Gasol	12.52449511	175.5	117	95.29	0	1.475	100	671.98
Carmelo Anthony	14.01846521	224	127	305.33	95.4	1	20	771.73
Damian Lillard	17.552592	182.5	200	195.18	95.2	1	20	692.88
Vince Carter	7.844822	213	76	174.13	58	1	0	521.13
Paul George	16.90243362	191	161	130.85	65.8	1	38	586.65
Derrick Rose	10.70573743	391.5	108	72.7	117.6	1	20	709.8
Rajon Rondo	7.368602313	90	53	163.17	47.2	1.285	81	535.08
Kevin Love	15.85205407	137	93	161.52	51.6	1.15	30	539.58
Klay Thompson	18.486885	114	55	64.48	24	1.675	147	578.29
LaMarcus Aldridge	13.4945385	155	122	144.65	66.2	1	0	487.85
Kyrie Irving	16.23522975	178.5	76	107.35	0	1.25	67	519.31
Blake Griffin	19.87450523	178.5	137	134.22	81.8	1	0	531.52
Draymond Green	11.90486736	122.5	92	57.75	31	1.405	102	528.07
Jimmy Butler	15.07803883	150	70	72.2	41.6	1	30	363.8
Marc Gasol	14.29212238	126.5	148	30.47	44	1.15	30	431.32
Kyle Lowry	12.67306147	155	25	79.41	24	1.15	48	373.92
Andre Iguodala	9.840313158	125	29	35.8	0	1.405	81	347.66
Nikola Jokic	14.42013638	63	146	51.68	60.8	1	20	341.48
DeMar DeRozan	14.47898357	139	70	119.15	45.6	1	20	393.75
DeAndre Jordan	10.8501128	25	144	161.18	0	1	0	330.18
Al Horford	15.2017645	139	39	43.35	0	1	0	221.35
John Wall	21.14585682	157	39	93.93	31	1	0	320.93
Rudy Gobert	13.8598196	125	100	94.04	22	1	0	341.04
Tyson Chandler	9.996318211	75	47	86.79	0	1.15	30	270.11
Joakim Noah	11.13622862	97.5	125	40.1	56	1	0	318.6
Joel Embiid	19.15347229	108.5	102	93.34	47	1	0	350.84
Paul Millsap	12.23898506	112	14	32.55	0	1	0	158.55
Isaiah Thomas	3.128064909	86.5	54	95.07	37	1	20	292.57
DeMarcus Cousins	8.355507	100.5	96	157.68	0	1	0	354.18
Kemba Walker	13.40152275	101	25	67.64	0	1	0	193.64
Serge Ibaka	9.931754429	0	36	64.04	0	1.1	20	130.05
Bradley Beal	16.16861009	68.5	10	95.88	0	1	0	174.38
Karl-Anthony Towns	14.26542938	97	33	89.38	0	1	0	219.38
Andre Drummond	12.72658836	62.5	29	180.75	0	1	0	272.25
Mike Conley	14.36325163	5	10	42.28	13	1	0	70.28
Brook Lopez	10.6119674	46	18	68.27	0	1	0	132.27
Goran Dragic	10.16416213	59	25	34.2	14	1	30	162.2
Ben Simmons	18.0367172	93.5	43	61.28	0	1	0	197.78
Pascal Siakam	9.995511	57	50	17.3	24	1.25	40	225.38
Luka Doncic	6.4935502	80	118	63.11	40	1	0	301.11

Player	mit_GMSC_re	mit_GMSC_pl	mit_WS_reg	mit_WS_pla	Commit_score	me_GMSC_reg	e_GMSC_pla	ime_WS_regu	me_WS_play	Prime_score	Legacy_regula	Legacy_playof	Legacy_final	Legacy_score	Total_score
LeBron James	200.14	241.45	206.31	270.17	918.07	199.71	147.97	205.6	176.37	729.65	154	354.4	147	655.4	10053.75
Kevin Durant	124.64	115.96	123.66	112.86	477.12	188.15	120	191.8	128.27	628.23	47	40	18	105	4531.66
Stephen Curry	93.16	87.7	90.06	83.54	354.46	179.55	112.88	166.32	117.05	575.8	66	40	36	142	4371.47
James Harden	115.76	89.15	117.81	81.59	404.32	207.78	116.19	166.32	81.77	572.06	30.5	0	0	30.5	2523.78
Kawhi Leonard	59	84.9	65.02	97.71	306.63	169.07	124.07	130.23	128.27	551.64	0	6.4	0	6.4	2463.32
Russell Westbrook	114.51	75.91	88.23	50.81	329.46	181.17	123.56	136.95	86.58	528.26	33	0	0	33	2304.56
Dwight Howard	116.6	63.59	117.12	64.49	361.79	150.52	105.76	146.51	102.62	505.41	8	0	15	23	2165.76
Chris Paul	133.03	79.35	157.35	80.61	450.34	172.83	117.46	184.72	83.38	558.39	48.5	0	0	48.5	2105.06
Giannis Antetokounmpo	63.97	32.99	56.2	28.34	181.5	191.65	113.9	136.95	59.33	501.82	0	0	0	0	2037.27
Anthony Davis	78.27	32.51	72.52	32.24	215.55	191.11	126.1	137.3	94.6	549.12	13	16	0	29	1929.09
Paul Gasol	129.31	74.92	125.76	75.73	405.72	141.39	91.27	143.67	137.89	514.22	6	0	0	6	1597.92
Carmelo Anthony	123.94	44.24	89.28	27.85	285.3	152.67	104.24	104.75	59.33	420.98	15	0	0	15	1493.01
Damian Lillard	79.33	36.36	71.47	21.01	208.17	173.91	93.05	128.46	46.5	441.92	5.5	0	0	5.5	1348.47
Vince Carter	127.4	43.29	109.35	36.15	316.18	150.26	107.8	124.56	62.53	445.15	27	0	0	27	1309.46
Paul George	67.38	50.08	61.09	43.48	222.03	145.42	109.32	112.89	73.76	441.38	0	0	0	0	1250.06
Derrick Rose	56.49	27.33	34.65	16.61	135.08	134.67	91.02	88.82	48.1	362.61	0	0	0	0	1207.49
Rajon Rondo	67.19	63.8	52.19	56.67	239.86	114.24	88.47	94.49	84.98	382.18	20	28	0	48	1205.12
Kevin Love	76.22	29.75	71.56	27.36	204.89	163.43	65.59	126.33	65.74	421.09	0	0	0	0	1165.56
Klay Thompson	55.66	58.56	40.84	42.5	197.57	119.34	74.24	84.58	70.55	348.71	20	16	0	36	1160.57
LaMarcus Aldridge	107.22	42.21	97.22	26.87	273.53	142.46	93.56	113.59	40.08	389.7	4	0	0	4	1155.08
Kyrie Irving	62.46	40.33	53.5	38.11	194.39	158.32	92.29	100.5	86.58	437.69	0	0	0	0	1151.39
Blake Griffin	74.47	33.58	65.63	24.92	198.59	145.69	99.41	115.36	56.12	416.57	0	0	0	0	1146.68
Draymond Green	39.82	64.37	38.75	68.4	211.34	107.79	77.03	98.38	94.6	377.8	0	0	0	0	1117.21
Jimmy Butler	61.7	47.55	65.54	47.88	222.66	158.32	98.14	120.67	89.79	466.92	0	0	0	0	1053.38
Marc Gasol	75.34	44.39	72	46.9	238.63	117.46	80.34	106.52	70.55	374.87	0	0	0	0	1044.82
Kyle Lowry	82.08	45.8	82.12	38.11	248.11	138.16	77.29	118.55	65.74	399.73	0	0	0	0	1021.76
Andre Iguodala	88.51	59.44	84.65	57.65	290.25	122.57	83.39	93.78	68.95	368.68	0	0	0	0	1006.59
Nikola Jokic	44.26	28.44	42.5	26.87	142.07	149.45	112.88	114.3	88.19	464.82	0	0	0	0	948.37
DeMar DeRozan	80.65	31.04	58.91	11.24	181.83	143.27	84.15	100.86	25.65	353.93	0	0	0	0	929.51
DeAndre Jordan	66.1	24	77.32	24.43	191.85	115.85	76.27	127.75	49.7	369.58	0	0	0	0	891.61
Al Horford	79.03	60.5	79.5	56.67	275.71	122.3	77.03	107.58	76.96	383.88	0	0	0	0	880.94
John Wall	63.71	25.77	38.66	16.61	144.75	139.77	110.85	86.7	35.27	372.59	10	16	0	26	864.27
Rudy Gobert	44.25	17	55.5	19.05	135.81	132.25	74.49	139.43	38.48	384.65	0	0	0	0	861.5
Tyson Chandler	71.92	21.53	89.1	31.27	213.83	97.84	58.22	105.46	65.74	327.25	0	0	0	0	811.19
Joakim Noah	46.93	25.62	54.02	27.85	154.42	106.71	70.68	97.32	44.89	319.6	0	0	0	0	792.62
Joel Embiid	27.56	15.02	19.9	10.26	72.74	154.29	95.34	73.96	25.65	349.24	0	0	0	0	772.82
Paul Millsap	83.76	48.16	80.03	38.11	250.06	118.27	89.75	92.72	41.69	342.42	0	0	0	0	751.03
Isaiah Thomas	48.27	14.85	39.53	7.33	109.98	140.58	78.81	105.81	20.84	346.05	0	0	0	0	748.6
DeMarcus Cousins	63.62	1.48	38.49	-0.49	103.1	158.86	0	76.08	0	234.94	0	0	0	0	692.22
Kemba Walker	69.72	16.09	47.47	11.73	145.01	139.77	77.29	94.13	35.27	346.47	0	0	0	0	685.12
Serge Ibaka	59.89	52.7	59.17	54.72	226.47	99.72	61.02	94.13	60.93	315.8	0	0	0	0	672.32
Bradley Beal	55.93	24.55	36.22	20.52	137.22	151.87	83.64	81.39	41.69	358.59	0	0	0	0	670.19
Karl-Anthony Towns	48.99	2.54	43.98	0.98	96.49	169.61	32.8	131.29	3.21	336.9	0	0	0	0	652.77
Andre Drummond	60.62	3.56	52.89	-0.49	116.57	134.4	0	106.87	0	241.27	8.5	0	0	8.5	638.59
Mike Conley	72.01	32.88	64.49	28.82	198.21	127.95	97.88	99.44	40.08	365.35	0	0	0	0	633.84
Brook Lopez	71.05	19.14	57.25	21.01	168.45	122.03	86.69	84.58	38.48	331.78	0	0	0	0	632.5
Goran Dragic	61.14	19.64	47.65	11.24	139.67	115.85	61.53	87.05	28.86	293.29	0	0	0	0	595.16
Ben Simmons	25.36	12.29	21.29	10.75	69.7	131.98	72.71	85.99	35.27	325.96	0	0	0	0	593.44
Pascal Siakam	20.06	20.34	18.24	15.63	74.28	98.38	64.58	68.65	44.89	276.5	0	0	0	0	576.16
Luka Doncic	18.01	5.52	11.96	2.44	37.93	103.75	59.49	48.48	8.02	219.74	0	0	0	0	558.78

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'.

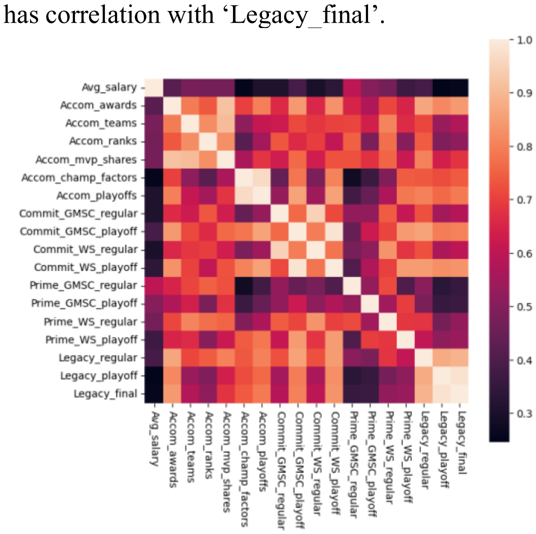

Figure 2: Correlation Matrix

3.2 Prediction

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

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

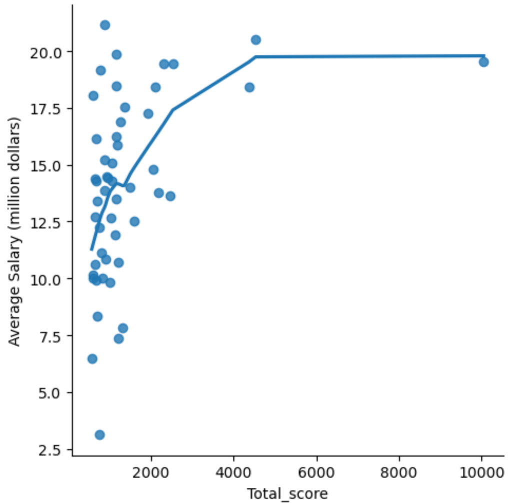

Figure 3: Simple Linear Regression results

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 [6], 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.

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 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:

	coef	std err	t	P>|t|
intercept	-13.51880	10.04800	-1.34500	0.18800
Accom_awards	0.00810	0.00900	0.91200	0.36900
Accom_teams	0.01290	0.01500	0.88500	0.38300
Accom_ranks	0.01050	0.01300	0.81800	0.41900
Accom_mvp_shares	-0.05950	0.03300	-1.78600	0.08400
Accom_champ_factors	13.38820	9.13500	1.46600	0.15300
Accom_playoffs	-0.09750	0.05400	-1.81900	0.07800
Commit_GMSC_regular	-0.11150	0.08900	-1.25600	0.21800
Commit_GMSC_playoff	0.38750	0.13700	2.82100	0.00800
Commit_WS_regular	0.00880	0.10400	0.08400	0.93300
Commit_WS_playoff	-0.31850	0.13200	-2.40600	0.02200
Prime_GMSC_regular	0.06820	0.03900	1.76100	0.08800
Prime_GMSC_playoff	0.03180	0.03500	0.89600	0.37700
Prime_WS_regular	0.02700	0.05400	0.50400	0.61800
Prime_WS_playoff	0.03900	0.04900	0.79600	0.43200
Legacy_regular	0.03600	0.07600	0.47400	0.63900
Legacy_playoff	-0.03470	0.05700	-0.61000	0.54600
Legacy_final	0.16320	0.14700	1.11000	0.27500

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 X_i; 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 = $\beta_0 + \beta_1$* 'Accom_awards' + β_2 * 'Accom_teams' + β_3* …. +β_4 * '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.

4. 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