ECON3389 Machine Learning in Economics

Course Project Report

Predictive Model of NBA Salaries

Group Number: 10

Group Members: Shyam Popat, Ashna Ramaswamy, Aman Sinha, Ishaan Masand, and Abhishek Rana

Contents

Introduction

In the United States, both professional and college sports are popular amongst the public. While the fun nature of these sports can be attributed to keeping the general public's interest, the contribution of each professional athlete to their respective sporting event is what has kept the competitive spirit of sports alive and relevant to this day. Regardless of the dominant sports culture in the US, there continue to be ongoing debates about the fairness of the compensation of pro athletes. The common denominator of this debate is the role that talent or performance plays in deciding a player's salary, regardless of the player's sex or participation in a specific sporting event.

In this report, we aim to explore the question of how much a player's talent or performance plays into the role of that specific player's salary. While there are many sports to consider for this study, we chose to look into the salaries of NBA athletes. The data on performance metrics and salaries for these players were taken from https://www.basketball-reference.com. This website includes information on the previously specified metrics of NBA players from official NBA matches across 18 seasons. Our proposed MLR model will consist of "salary (in dollars)" as the independent variable and performance metrics such as points per game, field goal percentage, assists, turnovers, age, and games as the independent variables. Since this regression will have more variables than simple linear regression, we are hoping for the regression to fit the data more accurately and provide us with a better look into the relationship between these variables and the salaries of NBA players. We believe that this model includes most of what can explain a player's salary as indicated by previous analyses completed by NBA statisticians. Moreover, our report will mention the details of the sample NBA dataset we used, the details of the inference and prediction models, and what we observe and how that helps to answer our main research questions.

Data Summary and Preliminary Analysis

All relevant graphs and data tables are attached in the appendix. Please refer to the specific descriptions below on where to find the summary statistics and key visualizations.

- See the attached table, titled “Table A1” in the appendix containing qualitative descriptions and units of measurements for the 39 variables included in this dataset.
- See the attached table, titled “Table A2” in the appendix containing the summary statistics for the 39 variables included in the dataset.
- See the attached graph, titled “Graph A3” in the appendix containing the key visualizations for the most important variables.

The noteworthy variables in this dataset were points per game, yearly salaries, minutes played, player ages, and field goal percentages. We believe these variables are important statistics used to measure a player’s value to the team, as well as overall performance. We found many of them to be significant in our linear regression model. Our visualizations show a positive correlation between player salaries and points per game, indicating that scoring is valued highly by teams. The visualization of player salaries yielded valuable insights as well. The salaries were severely right-skewed, meaning that the majority of players earn a low salary relative to the median, with around 4,500 players earning under 4 million dollars per year.

In comparison, fewer than 250 players earn over 20 million dollars a year. The highest density of players in the NBA is from ages 22-24, with the “age” histogram being severely right-skewed as well. The last pivotal visualization created was regarding field goal percentage. The field goal percentage histogram reflects a uniform distribution, as over 4000 players have a 40-50% field goal percentage. There are significantly smaller groups of players averaging field goal percentages outside of this range. This data summary effectively encompasses the primary and preliminary insights we wanted to draw from the dataset.

Inference and Predictive Models

We expected certain variables to be more indicative of player salary than others. Based on how players are usually evaluated, we predicted that these variables would include pts (points per game), fg.pct (field goal

percentage), ast (assists), tov (turnovers), ms (minutes played per game), age, and gs (games started in the season), orb (offensive rebounds), and drb (defensive rebounds). We ran a regression of these variables along with player salary to determine the correlation. After running the regression, we concluded that most of these variables were accurately predictive of player salary, as hypothesized. In particular, based on the coefficients, variables such as pts (coefficient of 521,646), drb (1,062,923), and age (339,871) had the most predictive power in determining player salary. These variables had a positive coefficient, which means they have a positive correlation with player salary in our linear regression. By contrast, the variable tov had a coefficient of -256,389, suggesting a negative correlation with player salary.

Next, we ran another regression to include the variables not included in our predictive model, to determine which variables were significant in predicting player salary. Based on the p-value of the regression, we determined that the following variables were significant: salary.lead (player's salary next year, dollars), age, g (games), gs (games started), fg.pct + efg.pct (effective field goal percentage) + fta (free throws attempted) + stl (steals) + blk (blocks) + tov (turnovers) + pf (personal fouls) + salary.team.mean (average salary of all players in each team for each year). Finally, we ran another regression using the listed variables to determine their coefficients and predictive power. Based on the coefficients of the variables in the regression, we learned that age (coefficient of 249,366), fg.pct (1,330,245), fta (178,814), and blk (435,648) have a positive correlation with player salary whereas pf (-193,449) has an inverse correlation with salary.

Ultimately, using the p-values from our regressions, we were able to narrow down the most significant variables in predicting player salary. We were also able to determine how and to what degree each of these variables affected salary. Doing so allowed us to conclude how teams value players and how certain statistics or other circumstances outside of their control can benefit or hurt their earnings.

Conclusion

We were able to establish several different variables and their respective polynomial terms that most accurately predicted a player’s salary in the NBA. Although we were able to obtain a relatively

accurate relationship between salary and the specific independent variables, a player's salary may not be exclusively based on the variables used in our model. Several non-quantifiable characteristics may increase or decrease a player's value to a potential franchise, such as a player's popularity. Our model offers valuable insights into quantifiable variables that have strong predictive power over a player's salary and can be used to gain a general understanding of the factors that may influence a player's salary, as well as the extent to which such factors influence earnings.

Appendix

Table A1: Qualitative description and units of measurements for the 39 variables included in this dataset.

variable	**description**	**unit of measurement**
playerid	Player ID code	qualitative
yrend	Year	year
teamid	Team	qualitative
salary	Player's salary	dollars
salary.lead	Player's salary next year	dollars
season	Season	year and month
teamfullsal	Team's Full name	qualitative
player	Player's Full Name	qualitative
position	Position	qualitative
age	Age of Player at the start of Feb 1st of that season	age
team	Team	qualitative
g	Games	no. of games
gs	Games Started	no. of games
mp	Minutes Played Per Game	min/game
fg	Field Goals Per Game	field goals/game
fga	Field Goals Attempts Per Game	field goals attempts/game
fg.pct	Field Goal Percentage	percentage of fgs
fg3	3 Point Field Goals Per Game	3-pt fgs/game

fg3a	3 Point Field Goal Attempts Per Game	3-pt fg attempts/game
fg3.pct	FG% on 3 Points Field Goal Attempts	percentage of 3-pt fgs
fg2	2 Point Field Goals Per Game	2-pt fgs/game
fg2a	2 Point Field Goal Attempts Per Game	2-pt fg attempts/game
fg2.pct	FG% on 2 Point Field Goal Attempts	percentage of 2-pt fgs
efg.pct	Effective Field Goal Percentage	percentage of fgs
ft	Free Throws Per Game	free throws/game
fta	Free Throws Attempts Per Game	free throws/game
ft.pct	Free Throw Percentage	percentage of free throws
orb	Offensive Rebounds Per Game	rebounds/game
drb	Defensive Rebounds Per Game	rebounds/game
trb	Total Rebounds Per Game	rebounds/game
ast	Assists Per Game	assists/game
stl	Steals Per Game	steals/game
blk	Blocks Per Game	blocks/game
tov	Turnovers Per Game	turnovers/game
pf	Personal Fouls Per Game	fouls per game
pts	Points Per Game	pts/game
salary.team.total	Total Salary of All of the players in each team for each year	dollars
salary.team.mean	Mean Salary of All of the players in each team for each year	dollars
salary.team.median	Median Salary of All of the players in each team for each year	dollars

Table A2: Summary Statistics of 39 variables in NBA Dataset

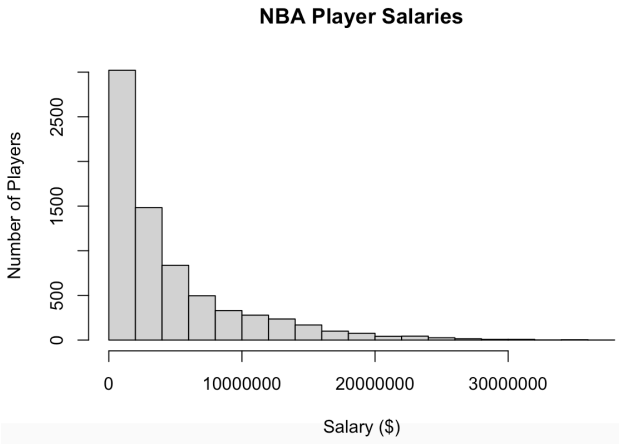
variable	**min**	**median**	**mean**	**max**
playerid	N/A	N/A	N/A	N/A
yrend	N/A	N/A	N/A	N/A
teamid	N/A	N/A	N/A	N/A
salary	4608	2626473	4653677	37457154

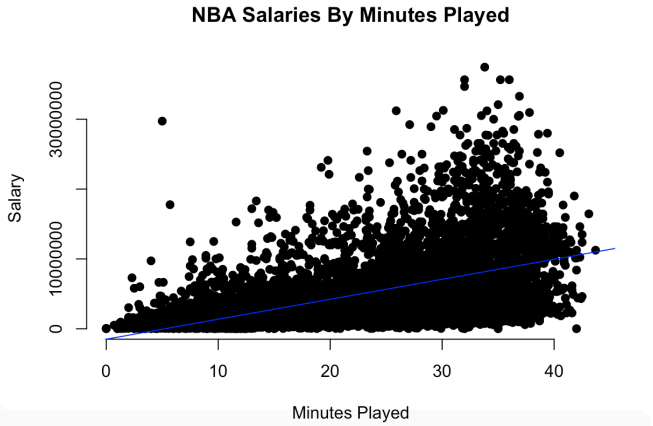
salary.lead	8819	3500000	5483953	37457154
season	N/A	N/A	N/A	N/A
teamfullsal	N/A	N/A	N/A	N/A
player	N/A	N/A	N/A	N/A
position	N/A	N/A	N/A	N/A
age	18	26	26.47	42
team	N/A	N/A	N/A	N/A
g	1	64	56.12	82
gs	0	15	28.62	82
mp	0	21.1	21.32	43.7
fg	0	2.8	3.245	12.2
fga	0	6.2	7.208	27.8
fg.pct	0	0.441	0.4422	1
fg3	0	0.3	0.6004	5.1
fg3a	0	1.1	1.705	13.2
fg3.pct	0	0.328	0.2823	1
fg2	0	2.1	2.644	11.2
fg2a	0	4.5	5.503	23.4
fg2.pct	0	0.472	0.4693	1
efg.pct	0	0.487	0.4809	1.5
ft	0	1.1	1.591	9.7
fta	0	1.6	2.118	13.1
ft.pct	0	0.753	0.7278	1
orb	0	0.7	1.002	5.5
drb	0	2.4	2.77	11.5
trb	0	3.2	3.769	16.3
ast	0	1.3	1.889	11.7
stl	0	0.6	0.6672	2.9
blk	0	0.3	0.442	3.7

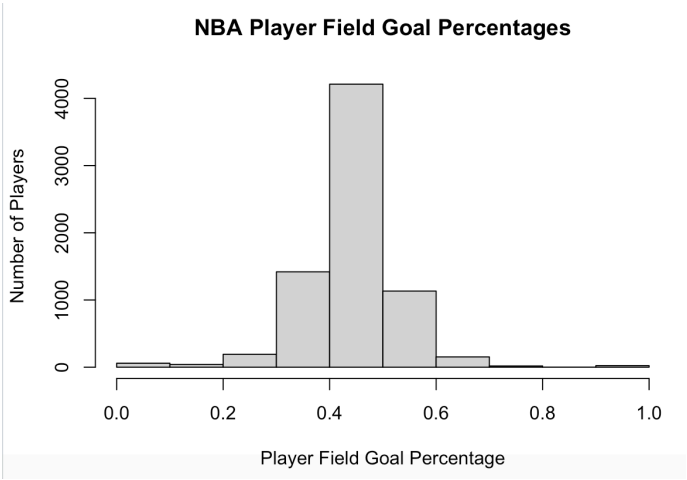
tov	0	1	1.225	5.7
pf	0	1.9	1.901	6
pts	0	7.2	8.679	36.1
salary.team.total	12806259	54981705	56659916	121749964
salary.team.mean	1601798	4248598	4399197	10145830
salary.team.median	789170	2650000	56659916	121749964

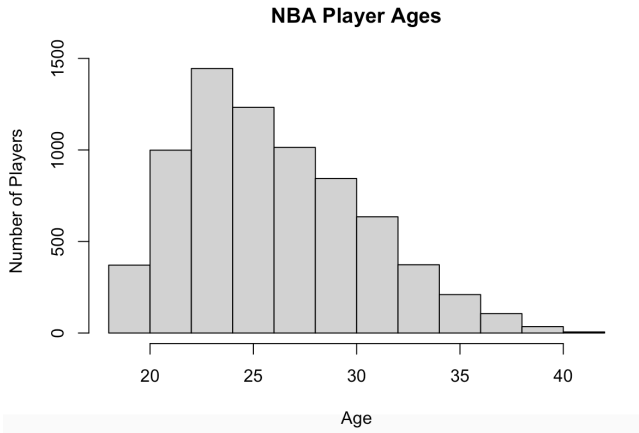
Note: Some of the variables, as mentioned previously, were qualitative and therefore do not have any values for the min/max, median, and mean. All such variables are marked with "N/A".

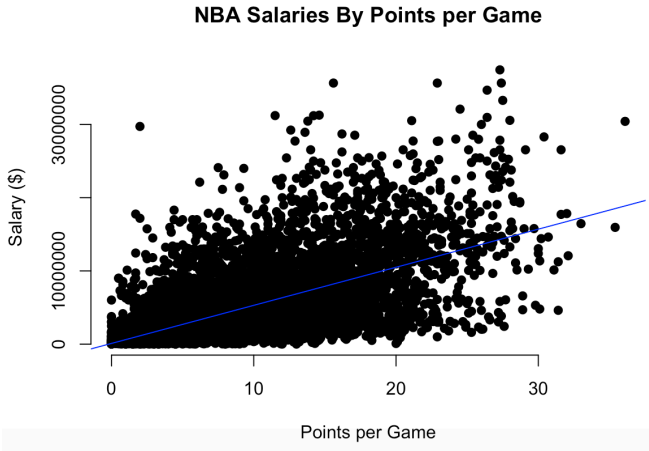
Graph A3: Key Visualizations for the Most Important Variables

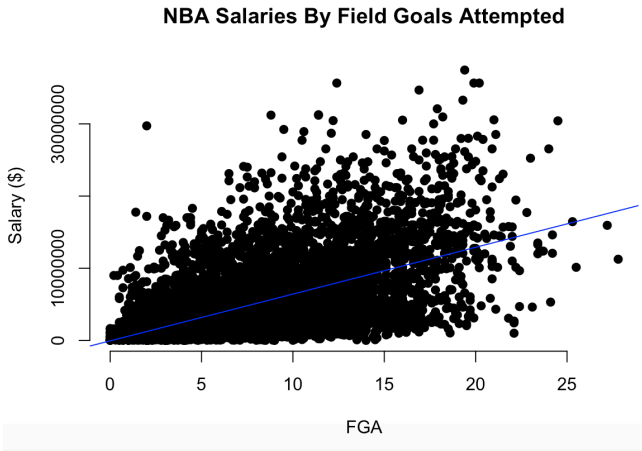

NBA Player Salaries
Number of Players
0
500
1500
2500
0
10000000
20000000
30000000
Salary ($)


NBA Salaries By Minutes Played
Salary
0
10000000
30000000
0
10
20
30
40
Minutes Played


NBA Player Field Goal Percentages
Number of Players
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0
Player Field Goal Percentage


NBA Player Ages
Number of Players
0
500
1000
1500
20
25
30
35
40
Age


NBA Salaries By Points per Game
Salary ($)
0
10000000
30000000
0
10
20
30
Points per Game


NBA Salaries By Field Goals Attempted
Salary ($)
0
10000000
30000000
0
5
10
15
20
25
FGA