My title*

My subtitle if needed

Yihang Cai

April 12, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/peachvegetable/NBA-player-points

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019a).

The remainder of this paper is structured as follows. Section 2….

2 Data

The dataset for this analysis was acquired from Basketball Reference Sports Reference LLC (2024) and includes a wide range of NBA player statistics for the 2023-2024 season. The process of downloading this dataset involved converting the website's data table into a CSV format, then transferring this data into Excel. In Excel, I employed the 'Text to Columns' feature to separate the statistics using commas, thereby preparing the dataset for analysis. This dataset comprises a variety of player statistics, such as position, age, assists, and steals, with a total of 718 observations before any data cleaning.

The analysis was conducted in the R statistical programming environment R Core Team (2023), utilizing a selection of packages for different tasks. Data cleaning was performed using the 'Tidyverse' Wickham et al. (2019b) and 'Janitor' Firke (2023) packages, while 'Dplyr' Wickham et al. (2023) and 'Broom' Robinson, Hayes, and Couch (2023) were used for data manipulation and data frame visualization. The 'Knitr' Xie (2014) and 'Ggplot2' Wickham (2016) packages were employed for data visualization, including the creation of tables and figures. Predictive modeling and model visualization were facilitated by the 'Tidymodels' Kuhn and Wickham (2020) and 'Modelsummary' Arel-Bundock (2022) packages, respectively.

The raw data is presented in Section A.1, divided into four separate tables (Table 7, Table 8, Table 9, Table 10). The dataset contains 30 variables, each thoroughly introduced and explained in Section A.1, offering a detailed view of the data that forms the basis of this study.

2.1 Data processing and interested predictors

This dataset comprises statistics for players, such as 3-point goals, 2-point goals, field goals, and free throws, from which points can be derived. My objective is to forecast an NBA player's total points based on their performance metrics, indicating a direct correlation between these features and the target variable, namely, points.

Table 1: Top 5 NBA Players Based on Select Predictors Highly Correlated with Points Scored, Season 2023-2024

Player	3-point goals	2-point goals	free throws	points
Precious Achiuwa	25	204	69	552
Precious Achiuwa	13	65	24	193
Precious Achiuwa	12	139	45	359
Bam Adebayo	10	470	276	1246
Ochai Agbaji	61	103	28	417

In Table 1, we see variables that are directly linked to a player's total points scored. For example, considering Bam Adebayo's performance in the 2023-2024 season: he scored 10 three-pointers, made 470 two-pointers, and successfully shot 276 free throws, totaling $10 \times 3 + 470 \times 2 + 276 \times 1 = 1246$ points, which exactly matches his recorded total points. To streamline the dataset for analysis, I utilized the 'tidyverse' package in R to remove these variables, which are 'fg', 'fga', 'x3p', 'x3pa', 'x2p', 'x2pa', 'ft', and 'fta'.

The dataset initially detailed players across 12 unique positions, which was quite detailed for modeling purposes. To simplify, I grouped these positions into three main categories: Guards (G): SG, PG, SG-PG, PG-SG, Forwards (F): SF, PF, PF-SF, and Centers (C): C, PF-C, C-PF. This grouping was intended to make the model clearer and to potentially improve its predictive power by reducing unnecessary complexity and avoiding overlap in variables.

Additionally, I re-evaluated the necessity of certain variables such as 'trb' (total rebounds), 'player', and 'tm' (team). 'trb', being the sum of 'orb' (offensive rebounds) and 'drb' (defensive rebounds), didn't provide additional insight and was thus omitted. The 'player' variable was removed in favor of 'rk' (rank), which sufficed for identifying players without duplicating information. Lastly, the 'tm' variable was excluded as the analysis didn't focus on team-specific performance, making the team data unnecessary for this study.

Table 2: Top 10 NBA Players with selectely statistics, Season 2023-2024

rk	age	g	gs	mp	fg%	x3p%	x2p%	efg%	ft%	orb	drb	ast	stl	blk	tov	pf	pts	pos
1	24	67	18	1522	0.51	0.27	0.57	0.54	0.62	184	277	94	44	66	78	130	552	C
1	24	25	0	437	0.46	0.28	0.53	0.50	0.57	50	86	44	16	12	29	40	193	C
1	24	42	18	1085	0.54	0.26	0.60	0.56	0.64	134	191	50	28	54	49	90	359	F
2	26	63	63	2162	0.52	0.33	0.53	0.53	0.75	142	529	253	73	61	148	144	1246	C
3	23	72	23	1457	0.41	0.30	0.53	0.49	0.67	66	128	73	42	38	55	102	417	G
3	23	51	10	1003	0.43	0.33	0.55	0.52	0.75	35	91	47	27	29	34	66	274	G
3	23	21	13	454	0.40	0.24	0.49	0.44	0.59	31	37	26	15	9	21	36	143	G
4	23	60	34	1595	0.44	0.35	0.53	0.53	0.62	72	277	136	43	51	69	87	652	F
5	25	74	19	1742	0.43	0.38	0.51	0.55	0.79	33	118	185	57	39	69	130	563	G

rk	age	g	gs	mp	fg%	x3p%	x2p%	efg%	ft%	orb	drb	ast	stl	blk	tov	pf	pts	pos
6	28	68	68	2284	0.50	0.47	0.57	0.66	0.88	43	218	213	60	42	87	144	915	G

As illustrated in Table 2, aside from 'rk' serving as an identifier, the selected variables are central to our analysis. These will act as predictors for estimating a player's total points, considering factors such as 3-point goal percentage, position, among others. Furthermore, for enhanced clarity, the values in the tables have been formatted to display two decimal places.

3 Model

This model pursues two main aims. The initial goal is to predict the total points an NBA player might score based on various performance indicators such as position and shooting efficiency. The second goal is to identify which predictors most significantly affect a player's scoring ability. For example, it is assumed that more playtime within a season could lead to a higher score.

To meet these objectives, the lasso regression model is chosen for its unique features: First, with 19 predictors left after processing the data, the lasso regression can reduce the influence of less important predictors by setting their coefficients to zero. Second, it clearly indicates which predictors have a greater impact on the scoring outcome, helping to understand what factors are most important in determining a player's points.

Lasso regression, a variant of linear regression models, is notable for its ability to select features by reducing the coefficients of less critical features to zero. This model introduces a regularization parameter, λ, which determines the strength of the penalty. This penalty minimizes some coefficients, especially those for less important variables, towards zero. As λ increases, more coefficients are reduced to zero, leading to a simpler model. The optimal value for λ is determined through cross-validation, ensuring the model is effectively tuned for the predictive tasks.

3.1 Model set-up

$$y_i = \beta_0 + \beta_i \cdot X_i \tag{1}$$

In this equation, y_i is the number of points a player scores, which is the dependent variable I am trying to predict. β_0 is the interception, and β_i is a matrix that contains the coefficients $\beta_1, \beta_2, ..., \beta_{18}$ for each predictor that the lasso regression will estimate. X_i is also a matrix contains the predictors: players position, age, games, game starts, minutes played, field goal percentage, 3-point field goal percentage, 2-point field goal percentage, effective field goal percentage, free throw percentage, offensive rebounds, defensive rebounds, total rebounds, assists, steals, blocks, turnovers, and personal fouls.

We run the model in R (R Core Team 2023) using the `tidymodels` package of Kuhn and Wickham (2020).

3.2 Model justification

We anticipate a positive correlation between the points scored and several factors: age, minutes played, number of games played, games started, shooting efficiency (encompassed by 2-point goal percentage, 3-point goal percentage, field goal percentage, and free throw percentage), rebounds (both offensive and defensive), and assists. The logic is straightforward: the higher these variables, the greater the likelihood of scoring more points. Additionally, a player's position could influence their scoring, as different positions in basketball have distinct objectives; for instance, center(C) may prioritize defense over scoring.

3.3 Model performance

Table 3: First lasso regression model top 10 predictions

Rank	Points	Prediction
3	274	372.78
7	1142	1128.62
13	247	337.65
19	265	263.01
26	56	29.30
29	1191	1275.16
30	1036	985.43
33	197	185.05
33	174	177.09
34	288	437.87

Table 3 displays the predictions made by the initial lasso regression model for the number of points scored by NBA players, numbered by 'Rank' identifier. For each 'Rank', there are two columns: 'Points', which represents the actual points scored, and 'Prediction', which shows the predicted points scored by the model. It's noticeable that there's a variance between the actual points and the predicted values. The model does not seem to accurately predict the points: In some cases, such as Rank 3, the model overestimates the points, predicting 372.78 points against the actual 274. In other instances, like Rank 7, the prediction is quite close to the actual points scored (1128.62 predicted vs. 1142 actual). There are also underestimations, as seen with Rank 30, where the model predicts 985.43 points while the actual points scored are 1036.

Table 4: RMSE and MAE of first lasso regression model

	RMSE	MAE
First lasso regression model	111.79	80.6

The table Table 4 lists two error metrics for assessing the first lasso regression model. RMSE(Root Mean Squared Error), recorded at 111.79, captures the average error by squaring the difference between the model's predictions and the actual points, thereby giving more weight to larger discrepancies and making it particularly useful where such errors have greater consequences. MAE(Mean Absolute Error), noted as 80.6, represents the simple average of all prediction errors without emphasis on their size, making it a reliable metric when treating all errors uniformly is preferable. Utilizing both RMSE and MAE provides a dual perspective: RMSE highlights the impact of substantial errors, and MAE offers a clear measure of average error, assisting in a balanced evaluation of the model's performance. This approach to error analysis suggests that the model's predictions could be improved by reevaluating the included features, especially in areas where the model's accuracy is critical.

3.3.1 Feature engineering

Table 5: Top 5 important variables of the first lasso regression model

Predictors	Coefficients
mp	310.86
tov	225.98
drb	73.62
orb	-70.80
pf	-65.66

Table 5 lists the predictors with the highest magnitude coefficients from the lasso regression model, indicating their relative importance in predicting the outcome variable. The listed predictors are minutes played ('mp'), turnovers ('tov'), defensive rebounds ('drb'), offensive rebounds ('orb'), and personal fouls ('pf'), with their corresponding coefficients.

The coefficient for 'mp' is positive (310.86), highlighting a direct relationship with point totals — more minutes played usually provides more opportunities for scoring. In contrast, 'orb' has a negative coefficient (-70.80), hinting at players with high offensive rebounds not necessarily correlating with higher points, perhaps indicating a focus on rebounding over scoring. 'Tov' carries a positive coefficient (225.98), which might seem counterintuitive given turnovers are adverse events; yet, it could reflect that players who handle the ball frequently might incur more turnovers and also have more scoring chances. A positive coefficient for ‘drb’ (73.62) suggests a link between securing defensive rebounds and higher point scores, likely due to the additional possessions gained. ‘Pf’ shows a negative coefficient (-65.66), indicating that fouling frequently could decrease a player’s scoring by reducing playing time due to foul trouble.

To refine the model, feature engineering introduced the ‘pts_per_min’ predictor, combining points with minutes played to assess scoring efficiency. This reflects how well players score relative to their time on the court. ‘tov_per_game’ adjusts turnovers for the number of games, enabling a fairer comparison across players, and ‘pf_per_game’ computes the average fouls per game, a significant aspect in evaluating defensive conduct and the potential impact on game participation and point contribution. These engineered features aim to provide a clearer understanding of each player’s performance, leading to an improved model with lower error metrics.

Table 6: RMSE and MAE of finalized lasso regression model

	RMSE	MAE
Second lasso regression model	94.68	66.14

As shown in Table 6, the second lasso regression model, enhanced with engineered features, has demonstrated significant improvement. The RMSE has decreased from 111.7 to 94.68, and the MAE has dropped from 80.6 to 66.14. This reduction in both metrics indicates that the model now predicts the number of points an NBA player could score based on their performances with greater precision.

4 Results

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

A.1 Raw data

raw data from basketball reference is split into four tables for a better view, which are displayed below

Table 7: Basic information and overall performance

rk	player	pos	age	tm	g	gs	mp	pts
1	Precious Achiuwa	PF-C	24	TOT	67	18	1522	552
1	Precious Achiuwa	C	24	TOR	25	0	437	193
1	Precious Achiuwa	PF	24	NYK	42	18	1085	359
2	Bam Adebayo	C	26	MIA	63	63	2162	1246
3	Ochai Agbaji	SG	23	TOT	72	23	1457	417

Table 8: Shooting efficiency

player	fg	fga	fg_percent	x3p	x3pa	x3p_percent	x2p	x2pa	x2p_percent	e_fg_percent
Precious Achiuwa	229	449	0.510	25	93	0.269	204	356	0.573	0.538
Precious Achiuwa	78	170	0.459	13	47	0.277	65	123	0.528	0.497
Precious Achiuwa	151	279	0.541	12	46	0.261	139	233	0.597	0.563
Bam Adebayo	480	922	0.521	10	30	0.333	470	892	0.527	0.526
Ochai Agbaji	164	396	0.414	61	200	0.305	103	196	0.526	0.491

Table 9: Free throws and rebounds

player	ft	fta	ft_percent	orb	drb	trb
Precious Achiuwa	69	112	0.616	184	277	461
Precious Achiuwa	24	42	0.571	50	86	136
Precious Achiuwa	45	70	0.643	134	191	325
Bam Adebayo	276	367	0.752	142	529	671

player	ft	fta	ft_percent	orb	drb	trb
Ochai Agbaji	28	42	0.667	66	128	194

Table 10: Playmaking and defence

player	ast	stl	blk	tov	pf
Precious Achiuwa	94	44	66	78	130
Precious Achiuwa	44	16	12	29	40
Precious Achiuwa	50	28	54	49	90
Bam Adebayo	253	73	61	148	144
Ochai Agbaji	73	42	38	55	102

1. rk: rank - this doesn't represent the ranking of players based on some criterion, but purely for numbering purpose
2. player: player - the name of the basketball player.
3. pos: position - the playing position of the player.
4. age: the age of each player.
5. tm: team - the abbreviation of the NBA team the player belongs to.
6. g: games - how many games a player played in this season.
7. gs: game started - how many games a player has been in the starting lineup for their team at the beginning of the game.
8. mp: minutes played - the total time of a player played in this season.
9. fg: field goals - the total number of field goals (baskets) the player has made.
10. fga: field goal attempts - the total number of field goal shots the player has attempted.
11. fg_percent: field goal percentage - this statistic represents the percentage of field goals (both 2-pointers and 3-pointers) made by a player out of the total number of field goal attempts.
12. x3p: 3-point field goals - the total number of 3-point field goals the player has made.
13. x3pa: 3-point field goal attempts - the total number of 3-point shots the player has attempted.
14. x3p_percent: 3-point goal percentage - this statistic represents the percentage of 3-point field goals made by a player out of the total number of 3-point field goal attempts.
15. x2p: 2-point field goals - the total number of 2-point field goals the player has made.

16. x2pa: 2-point field goal attempts - the total number of 2-point shots the player has attempted.
17. x2p_percent: 2-point goal percentage - this statistic represents the percentage of 2-point field goals made by a player out of the total number of 2-point field goal attempts.
18. e_fg_percent: effective field goal percentage - this statistic adjusts for the fact that a 3-point field goal is worth more than a 2-point field goal.
19. ft: free throws - the total number of free throws the player has made.

20: fta: free throw attempts - the total number of free throw shots the player has attempted.

21. ft_percent: free throw percentage - this statistic represents the percentage of free throws made by a player out of the total number of free throw attempts.
22. orb: offensive rebounds - this statistic represents the number of rebounds grabbed by a player on the offensive end of the court.
23. drb: defensive rebounds - this statistic represents the number of rebounds grabbed by a player on the defensive end of the court.
24. trb: total rebounds - this statistic represents the total number of rebounds grabbed by a player (both offensive and defensive rebounds).
25. ast: assists - the total number of assists the player has made, indicating the number of times a player's pass led directly to a basket by a teammate.
26. stl: steals - the total number of times the player has taken the ball away from an opponent, leading to a change in possession.
27. blk: blocks - the total number of times the player has deflected an opponent's filed goal attempt, preventing the ball from going into the basket.
28. tov: turnovers - the total number of times the player has lost possession of the ball to the opposing team.
29. pf: personal fouls - the total number of personal fouls the player has committed.

30: pts: points - the total number of points the player has scored.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Sports Reference LLC. 2024. "2023-2024 NBA Player Stats: Totals." Basketball-Reference.com. https://www.basketball-reference.com/leagues/NBA_2024_totals.html.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.