My title*

My subtitle if needed

Yihang Cai

April 8, 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. (2019).

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

2 Data

The dataset is obtained from Basketball Reference Sports Reference LLC (2024), which contains NBA players' statistics in the 2023 to 2024 season. The dataset is directly downloaded from Basketball Reference using the given instruction. The dataset contains different statistics of NBA players, for instance, players' position, age, assists, steals, etc. There are in total of 718 observations before any data cleanings.

Raw data is displayed in Section A.1, which is split into four different tables Table 3, Table 4, Table 5, Table 6. There are 30 variables in the dataset and each is introduced and explained in Section A.1.

2.1 Data processing and interested predictors

This dataset contains player statistics including 3-point goals, 2-point goals, field goals, and free throws, points can be calculated using these variables. My goal is to predict the number of points a NBA player scores based on his performance, which means that these features are too closely related to the target variable (i.e. 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
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 Table 2 shows, apart from 'rk' for identification purpose, these variables are what we interested in and will be used as predictors to predict the total points a player would potentially gain based on his performances like 3-point goal percentage, position and etc.

3 Model

Lasso regression is a variation of linear regressions that can actually perform feature selection by setting the coefficients of less important features to zero. It has an additional parameter λ, which is the regularization parameter that controls the strength of the penalty. The penalty term shrinks some of the coefficients (i.e. the less important variables) toward zero. As λ increases, more coefficients are set to zero, leading to a simpler model. The value of λ is determined through cross-validation.

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 expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

```
# #| include: false
# #| warning: false
# #| message: false
#
# first_lasso_model <- readRDS("../models/first_lasso_model.rds")
# second_lasso_model <- readRDS("../models/second_lasso_model.rds")
# predictions <- predict(second_lasso_model, test_data)
# results <- bind_cols(test_data, predictions)
# # Calculate RMSE
# rmse_results_2 <- rmse(results, truth = pts, estimate = .pred)
# rsq_results_2 <- rsq(results, truth = pts, estimate = .pred)
```

4 Results

Our results are summarized in **?@tbl-modelresults**.

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 3: 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 4: 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 5: 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 6: 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

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/.

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