# More Minutes, More Points — The Role of Aggressive Play in NBA Scoring Trends*

### An Analysis of NBA Player Performance Using Lasso Regression for the 2023-2024 Season

Yihang Cai

April 19, 2024

This paper investigates the predictive capacity of lasso regression in determining NBA players' total points from season statistics. The study found that minutes played, turnovers, and personal fouls are significant predictors of scoring, with turnovers unexpectedly indicating a higher involvement in gameplay leading to increased scoring. The implications of these findings extend beyond sports analytics, suggesting that player evaluations in team sports may benefit from a deeper understanding of player involvement as opposed to conventional performance metrics. By refining predictive models and reassessing performance indicators, this research offers new insights into effective player performance assessment and strategic planning in professional basketball.

## Table of contents

**1 Introduction** **2**

**2 Data** **3**

2.1 Data processing and interested predictors . . . . . 3

**3 Model** **5**

3.1 Model set-up . . . . . 6

3.2 Model justification . . . . . 6

3.3 Model performance . . . . . 6

3.3.1 Feature engineering . . . . . 7

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

**4 Results** **9**
4.1 Model overview . . . . . 9
4.2 Important predictors . . . . . 11

**5 Discussion** **12**
5.1 Model summary and modifications . . . . . 12
5.2 Key Predictive Variables . . . . . 12
5.3 Weaknesses . . . . . 13
5.4 Next steps . . . . . 14

**Appendix** **15**

**A Additional data details** **15**
A.1 Raw data . . . . . 15

**B References** **18**

# 1 Introduction

Professional basketball is not just a display of athletic skill and competition but also a valuable area for data analysis, especially in predicting player performances. This paper examines NBA player statistics from the 2023-2024 season, aiming to predict players' total points using various performance metrics. The study utilizes lasso regression, a statistical method appreciated for its feature selection and regularization capabilities, to enhance the accuracy and simplicity of our predictive model.

This research addresses the need for a predictive model that accurately forecasts player points while demonstrating the effects of different performance metrics on scoring. The main objective is to develop a reliable model that can quantify how specific metrics influence the total points scored by NBA players.

In pursuit of this goal, the paper investigates several performance indicators such as minutes played, turnovers, rebounds, and personal fouls. Through feature engineering, these metrics are optimized for their relevance in predicting player scores. The study details the rationale behind selecting lasso regression, which reduces the influence of less critical metrics, thereby focusing on the most impactful factors.

The findings highlight that common metrics like minutes played are significant predictors of scoring potential. Interestingly, it also shows that turnovers have a positive correlation with scoring, suggesting that players involved in aggressive offensive plays tend to score higher, despite the potential risks.

These insights are vital for team strategies and player assessments and broaden the application of statistical models in sports science. They offer both theoretical and applied insights into the complex dynamics of basketball performance.

The estimand of this study, the total points a player scores, is predicted based on various processed performance metrics using lasso regression. This focus is importnat to achieving the study's goals of enhancing predictive precision and understanding how different metrics influence scoring in NBA games.

The paper is structured as follows: Section 2 covers the data collection and preparation methods; Section 3 explains the model setup, justification, and evaluation of its performance; Section 4 discusses the results; and Section 5 interprets these results in detail. Appendix provides additional data details, rounding out the discussion and ensuring thorough documentation of all methods and analyses used.

# 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' by Wickham et al. (2019) and 'Janitor' by 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 was done by the 'Tidymodels' Kuhn and Wickham (2020).

The raw data is presented in Section A.1, divided into four separate tables (Table 8, Table 9, Table 10, Table 11). 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, which can be used to derive the number of points scored directly. Since

my objective is to forecast an NBA player's total points based on their performance metrics, variables that are too closely related to the target variable, namely points should be removed.

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', 'rk' (rank), and 'tm' (team). 'trb', being the sum of 'orb' (offensive rebounds) and 'drb' (defensive rebounds), didn't provide additional insight and was thus omitted. I also replaced 'player' and 'id' with a new variable 'id', which is sufficient for uniquely identifying players without redundancy. The 'rk' variable often duplicated because players sometimes transfer between teams within a season, making 'rk' and 'player' unreliable identifiers. Instead, I introduced 'id', assigned by row number, to effectively differentiate players across different teams, ensuring each instance is treated distinctly. Therefore, the same person in a different team would have a different id. 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

| id | age | g | gs | mp | fg% | x3p% | x2p% | efg% | ft% | orb | drb | ast | stl | blk | tov | pf | pts | pos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 24 | 67 | 18 | 1522 | 1 | 0 | 1 | 1 | 1 | 184 | 277 | 94 | 44 | 66 | 78 | 130 | 552 | C |
| 2 | 24 | 25 | 0 | 437 | 0 | 0 | 1 | 0 | 1 | 50 | 86 | 44 | 16 | 12 | 29 | 40 | 193 | C |

| id | age | g | gs | mp | fg% | x3p% | x2p% | efg% | ft% | orb | drb | ast | stl | blk | tov | pf | pts | pos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 24 | 42 | 18 | 1085 | 1 | 0 | 1 | 1 | 1 | 134 | 191 | 50 | 28 | 54 | 49 | 90 | 359 | F |
| 4 | 26 | 63 | 63 | 2162 | 1 | 0 | 1 | 1 | 1 | 142 | 529 | 253 | 73 | 61 | 148 | 144 | 1246 | C |
| 5 | 23 | 72 | 23 | 1457 | 0 | 0 | 1 | 0 | 1 | 66 | 128 | 73 | 42 | 38 | 55 | 102 | 417 | G |
| 6 | 23 | 51 | 10 | 1003 | 0 | 0 | 1 | 1 | 1 | 35 | 91 | 47 | 27 | 29 | 34 | 66 | 274 | G |
| 7 | 23 | 21 | 13 | 454 | 0 | 0 | 0 | 0 | 1 | 31 | 37 | 26 | 15 | 9 | 21 | 36 | 143 | G |
| 8 | 23 | 60 | 34 | 1595 | 0 | 0 | 1 | 1 | 1 | 72 | 277 | 136 | 43 | 51 | 69 | 87 | 652 | F |
| 9 | 25 | 74 | 19 | 1742 | 0 | 0 | 1 | 1 | 1 | 33 | 118 | 185 | 57 | 39 | 69 | 130 | 563 | G |
| 10 | 28 | 68 | 68 | 2284 | 1 | 0 | 1 | 1 | 1 | 43 | 218 | 213 | 60 | 42 | 87 | 144 | 915 | G |

As illustrated in Table 2, aside from 'id' 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 only integers, since it is unlikely to have 0.3 point in scoring in basketball.

## 3 Model

This model pursues two main goals. 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, $\lambda$, which determines the strength of the penalty. This penalty minimizes some coefficients, especially those for less important variables, towards zero. As $\lambda$ increases, more coefficients are reduced to zero, leading to a simpler model. The optimal value for $\lambda$ 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. $\beta_0$ is the interception, which represents the average value of y when x is 0, and $\beta_i$ 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, assists, steals, blocks, turnovers, and personal fouls. ID is not included here since it's just an identifier.

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: 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. A negative correlation may appear between age and the number of points scored, since it would be natural to assume that an older player could be weaker than a younger player due to aging, and thus scores less.

### 3.3 Model performance

Table 3: First lasso regression model top 10 predictions

| ID | Points | Prediction |
|---|---|---|
| 1 | 552 | 585 |
| 2 | 193 | 184 |
| 3 | 359 | 413 |
| 9 | 563 | 701 |
| 15 | 26 | -1 |
| 18 | 408 | 453 |
| 35 | 921 | 907 |

| ID | Points | Prediction |
|---|---|---|
| 44 | 23 | 37 |
| 53 | 179 | 143 |
| 56 | 114 | 132 |

Table 3 displays the predictions made by the initial lasso regression model for the number of points scored by NBA players, numbered by 'ID' identifier. For each 'ID', 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 ID 9, the model overestimates the points, predicting 701 points against the actual 563. In other instances, like ID 2, the prediction is quite close to the actual points scored (184 predicted vs. 193 actual). There are also unrealistic estimations, as seen with ID 15, where the model predicts -1 points while the actual points scored are 26, which it is impossible to have negative points scored in basketball.

Table 4: RMSE and MAE of first lasso regression model

| | RMSE | MAE |
|---|---|---|
| First lasso regression model | 110.12 | 78.23 |

Table 4 lists two error metrics for assessing the first lasso regression model. RMSE(Root Mean Squared Error), recorded at 110.12, 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 78.23, 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 re-evaluating 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 | 305 |
| tov | 209 |
| drb | 71 |
| orb | -60 |
| pf | -58 |

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 (305), highlighting a direct relationship with point totals — more minutes played usually provides more opportunities for scoring. In contrast, 'orb' has a negative coefficient (-60), 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 (209), 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' (71) suggests a link between securing defensive rebounds and higher point scores, likely due to the additional possessions gained. 'pf' shows a negative coefficient (-58), 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.

To guarantee that the final lasso regression model's predictions stay within realistic limits, we've implemented a simple yet deliberate adjustment: all negative predicted values are set to zero. While this approach may appear overly simplistic, it is employed for considered reasons that will be elaborated upon in the Discussion section (Section 5).

These engineered features aim to provide a clearer understanding of each player's performance, leading to an improved model with lower error metrics.

# 4 Results

## 4.1 Model overview

Table 6: RMSE and MAE of finalized lasso regression model

| | RMSE | MAE |
|---|---|---|
| Second lasso regression model | 91.03 | 60.8 |

As shown in Table 6, the second lasso regression model, enhanced with engineered features, has demonstrated significant improvement. The RMSE has decreased from 110.12 to 91.03, and the MAE has dropped from 78.23 to 60.8. 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.

Table 7: Final lasso regression model top 10 predictions

| ID | Points | Prediction |
|---|---|---|
| 1 | 552 | 558 |
| 2 | 193 | 209 |
| 3 | 359 | 342 |
| 9 | 563 | 622 |
| 15 | 26 | 0 |
| 18 | 408 | 418 |
| 35 | 921 | 921 |
| 44 | 23 | 0 |
| 53 | 179 | 219 |
| 56 | 114 | 107 |

Table 7 presents the comparison of actual points scored by players against the points predicted by our refined lasso regression model. For each player, designated by a ID, the table lists both the actual and predicted points, showcasing the model's predictions. For example, a player at ID 1 is shown with an actual score of 552 and a closely aligned prediction of 558. Similarly, for a player at ID 56 with an actual score of 114, the model predicts a nearly precise 107 points, and another at ID 3 has an actual score of 359 with a prediction of 342, reflecting the model's accuracy.

When we compare this to the predictions in Table 3, we notice a substantial improvement. Previously, for a player ID 9, the model had overestimated with a prediction of 701 points, which was quite higher than the actual, and for a player ID 15 the model had an unrealistic

estimation of -1. In contrast, Table 7 shows a more accurate and realistic prediction, indicating that the model has been better tuned.

However, in Table 7, there are instances where predictions have been adjusted to 0, such as at ID 15 and 44. This adjustment, while necessary to avoid negative predictions, could introduce bias. Further explanation and exploration of this methodological choice will be provided in Section 5 of the paper.

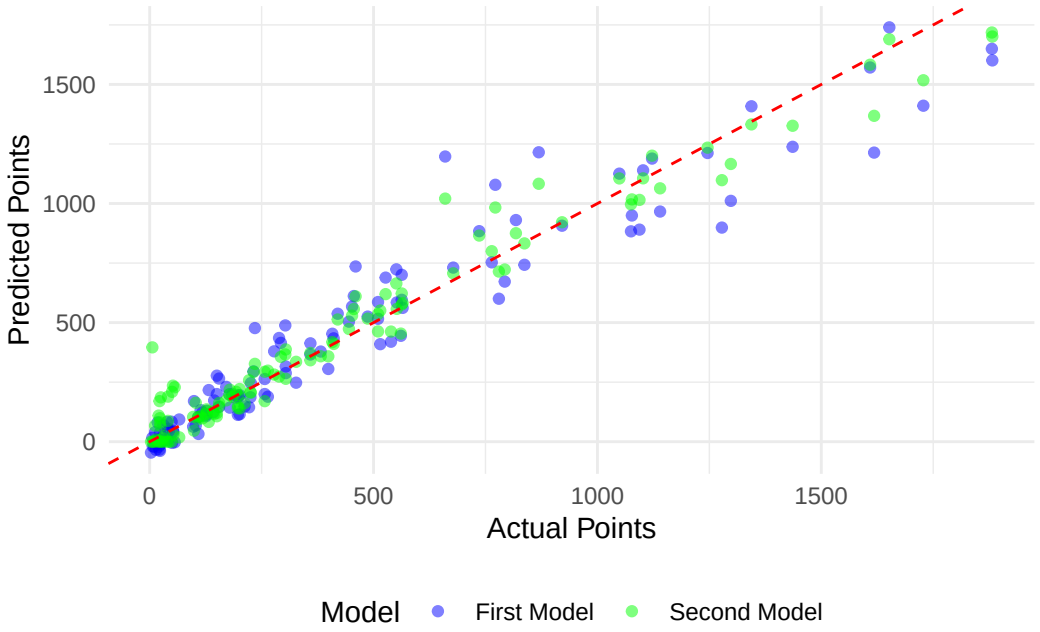
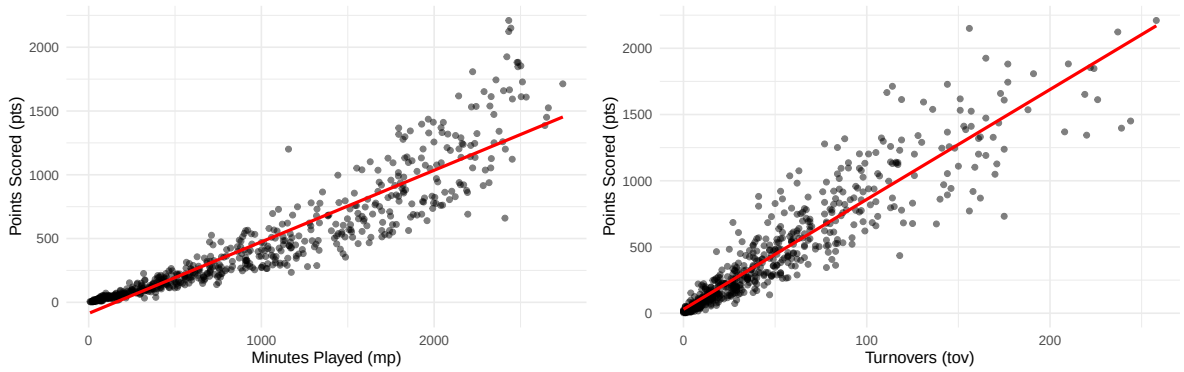



Figure 1: Comparison of first model prediction and second model prediction

Figure 1 is a scatter plot, which illustrates a comparison between actual points scored by players and predictions made by two different lasso regression models. The x-axis denotes the actual points, while the y-axis corresponds to the predicted points from each model. The green dots, representing predictions from the second model, are observed to cluster more closely to the red dashed line, which signifies perfect accuracy where predicted points match actual points. This clustering indicates that the second model generally has better predictive accuracy compared to the first model, whose predictions are denoted by the other color, perhaps blue, and may not cluster as tightly around this line of accuracy. However, it can be seen from the figure that the predictions made by the second model are further from the actual values compare with the predictions made by the first model. The possible reason will be discussed in Section 5.

## 4.2 Important predictors

According to Table 5, predictors 'mp' and 'tov' are the most important variables that contribute to the prediction of number of points as indicated by their coefficients in the model (305 and 209 respectively, much higher than the other three variables). To explore how these variables correlate with the scoring outcome, scatter plots were made to display their relationships with the target variable number of points.




(a) minutes played vs points

(b) turnovers vs points

Figure 2: Impact of player performance metrics on scoring across key statistical relationships

Figure 2 illustrates the relationships between various performance metrics and the total points scored by players, with each plot featuring a distinct statistical category. The linear trend lines, maided in red, offer a visual summary of each relationship. Figure 2a shows a positive correlation between the minutes played ('mp') and the points scored ('pts'). The upward trend suggests that players who spend more time on the court tend to score more points. This is intuitive and proves our assumption as more playing time offers more opportunities to score. Figure 2b displays the relationship between turnovers ('tov') and points scored also appears to be positive. Typically, turnovers are considered negative events; however, this positive trend may imply that players who are more involved in action-packed facets of the game tend to score higher, even though they might also commit more turnovers.

The importance of these variables in predicting the number of points scored is underpinned by their direct and indirect contributions to a player's impact on the game. Minutes played directly increases opportunities for scoring, and turnovers may indicate a player's involvement in high-risk, high-reward plays. These findings support the decision to include these variables in the predictive model because they play a significant role in predicting the number of points scored.

# 5 Discussion

The ability to predict player performance metrics in sports is a critical aspect of team management and strategy development. In this paper, we've created a model that accurately forecasts NBA players' total points using a variety of performance metrics.

## 5.1 Model summary and modifications

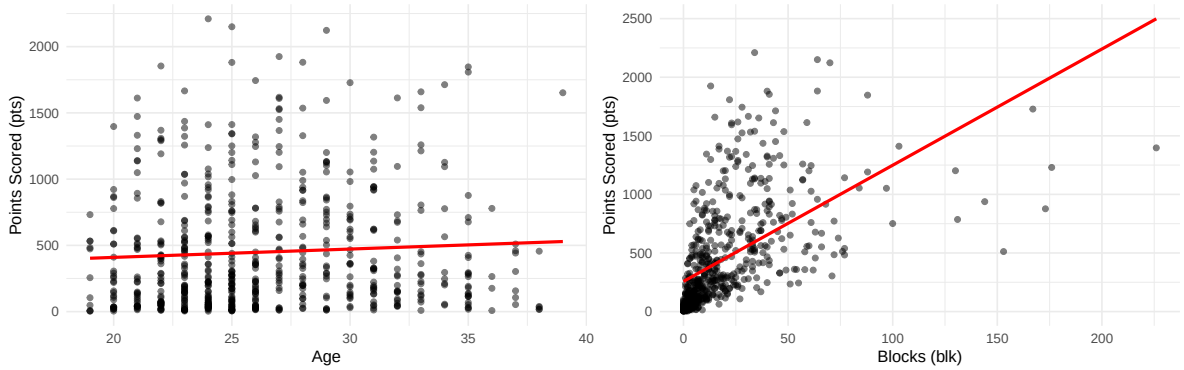
In this paper, we have applied a lasso regression model, a recognized method in statistics, to estimate the total points scored by NBA players from their performance statistics. This approach was selected for its ability to handle numerous predictors and its efficiency in shrinking less relevant predictors' coefficients towards zero, thus simplifying the model.

We refined the model through feature engineering, a significant step in which we derived new variables such as points per minute played, turnovers per game, and personal fouls per game. These variables were specifically constructed to capture the efficiency and impact of a player's performance relative to their time on the court and their involvement in the game. Moreover, we addressed the issue of negative predictions—impossible in the real-world context of point scoring—by setting a floor value of zero. While these modifications led to improvements in predictive accuracy, they also introduced certain limitations that we'll discuss subsequently.

## 5.2 Key Predictive Variables

Our findings reveal that minutes played, turnovers, total rebounds, and personal fouls hold substantial predictive power for determining a player's total points. These variables contribute significantly to a player's performance score, with minutes played emerging as a particularly potent predictor as shown in Section 4. This aligns with the intuitive understanding that more time on the court allows for more opportunities to score. However, apart from the key predictors we've discussed through the whole paper, let's explore those less important variables.

Similar with Figure 2, Figure 3 provides scatter plots that illustrate the relationships of age and blocks with points scored. I made a naive assumption that older players may score less than younger players, since in basketball, the physical demands on players can lead to a decline in performance as they age. Older players may score fewer points compared to their younger counterparts, primarily due to the cumulative effects of wear and tear on their bodies from multiple seasons of intense competition. This decline is often exacerbated by injuries that become more prevalent and harder to recover from with age. As players experience more seasons, the stress of high-intensity games can lead to persistent health issues, which may significantly impact their ability to maintain peak performance levels, particularly in scoring. However, Figure 3a shows that there isn't a noticeable trend between age and points scored, which contradicts with my assumption. This could be due to the following reasons: first, as




(a) age vs points

(b) blocks vs points

Figure 3: Less important performance metrics on points prediction

players age, they often adapt their playing style to stay competitive. Veterans might shift to roles that require less physical exertion or focus on aspects of the game that rely more on experience and skill rather than physicality. Secondly, modern training methods, diet, and medical treatments can prolong a player's peak performance. Players might maintain a high level of play longer than in the past due to better overall health and fitness management. Lastly, the most important factor could be that the cumulative knowledge and skills acquired from years of playing can help older players adapt and find new ways to score, effectively balancing out the physical challenges of aging.

Moreover, Figure 3b shows that the relationship between the number of blocks and points scored is weak, indicating that some players have roles that focus more on defense and assists rather than scoring.

## 5.3 Weaknesses

The study's methodology presents several weaknesses. Firstly, the potential non-linear relationships between some predictors and the total points scored are not fully captured by the lasso regression model. This could lead to an oversimplification of the complex dynamics of basketball performance. Secondly, the method of setting negative predictions to zero, although it ensures realistic prediction values, it might reduce the model's accuracy as it deviates from true statistical adjustments, such as those achievable through log transformation. However, log transformation brings more uncertainties to the prediction as applying the logarithm function to points can modify how points relate to other predictors, which in turn might decrease the precision of the model's predictions. Thirdly, making a more general grouping of positions can simplify the model and avoid making unnecessary dummy variables, but it also risks oversimplifying the distinct roles of each position, which could improve the model's performance if considered more carefully. Lastly, leaving out the 'tm' (team) variable fails to account for

how changes in team context might influence a player's performance, potentially skewing the model's predictive power.

### 5.4 Next steps

To build upon this study's foundation, future research should consider incorporating data from multiple seasons to provide a stronger training set for the model, potentially capturing more subtle patterns in player performance over time. Including the 'tm' variable in the analysis might offer insights into how team dynamics influence individual performance. Further exploration into non-linear modeling techniques could also better accommodate the complex relationships within the data.

# 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 8: 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 9: 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 10: 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 11: 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. id: 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.

## B References

weaknesses: position grouping; discussion: less important variables, like blk not served as scorer, age

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. 2019. "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.