

**ST. XAVIER'S COLLEGE (AUTONOMOUS) , KOLKATA**  
**DEPARTMENT OF STATISTICS**



**PREDICTING NBA ALL-STARS: A DATA-  
DRIVEN APPROACH USING PCA AND  
LOGISTIC REGRESSION**

**NAME: SOUMYA KARMAKAR**

**ROLL: 444**

**REGISTRATION NUMBER : A01-1142-0720-22**

**SEMESTER: 6**

**SESSION : 2022-2025**

**SUPERVISOR : DR. AYAN CHANDRA**

Declaration: I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Signature: \_\_\_\_\_

## Acknowledgment

I would like to express my gratitude to my mentor Dr. Ayan Chandra whose guidance and experience have been instrumental in the completion of this project.

I would also like to express my true gratitude to the department faculty members' valuable advice and continuous guidance that inspired me to continue improving myself and finally complete my project. Their expertise is unrivaled.

I am also grateful for the support and the good environment my family and friends facilitated throughout the process of me completing my project. The sharing and exchange of ideas and opinions helped me make my work so much better.

Lastly but not the least, I would like to thank my parents for their unconditional support, particularly during difficult times.

## Table of Contents

| Topics   | Page |
|--|------|
| 1) Introduction.....   | 4    |
| 2) Objectives.....   | 5    |
| 3) Description of dataset.....                                     | 6    |
| 4) Sources of dataset.....   | 7    |
| 5) EDA.....  | 8    |
| 6) Principal Component Analysis.....                               | 16   |
| Computation of first principal component.....                      | 16   |
| Computation of second principal component.....                     | 17   |
| Generalization of first M principal component.....                 | 17   |
| Insights from PCA on NBA data.....                                 | 18   |
| Scree Plot.....  | 19   |
| 7) Logistic Regression.....  | 21   |
| Roc Curve and other performance matrices.....                      | 24   |
| Deviance.....  | 26   |
| 8) Comparing Predicted vs Actual 2025 NBA All Star Selections..... | 27   |
| Observations.....  | 27   |
| Note on NBA All Star Games 2025.....                               | 28   |
| 9) Conclusion.....   | 28   |
| 10) R Codes.....   | 29   |
| 11) References.....  | 29   |

## Introduction

Player performance, team strategy, and team, coach, and analyst decision-making have all been revolutionized by sports analytics. Improved data availability and computation capabilities have made analytics a part of the life of sports, including basketball, soccer, and baseball. Statistical models complement traditional scouting in basketball with data on player efficiency, team building, and game planning.

Statistical metrics like Player Efficiency Rating (PER), Box Plus/Minus (BPM), and True Shooting Percentage (TS%) help quantify a player's contribution above and beyond traditional box score statistics. One of the most exciting applications of sports analytics is predictive modelling, which applies machine learning and statistical models to forecast game outcomes, player performance, and awards.

The **National Basketball Association (NBA)** is one of the most prestigious and competitive professional basketball leagues in the world. Established in 1946, the league has grown into a global sporting powerhouse, featuring elite athletes, thrilling matchups, and a rich history of legendary players. The NBA is known for its high-paced gameplay, advanced analytics, and continuous evolution in strategy and performance measurement.

One of the most anticipated events in the NBA calendar is the **NBA All-Star Game**, an annual exhibition showcasing the league's top players. Selection for this event is based on a combination of **fan votes, media opinions, and coach selections**, making it both a prestigious honor and a subject of debate. Popularity, team success, and previous All-Star appearances often influence the selection process, raising an important question:

**Is NBA All-Star selection possible using only player performance statistics?**

**The goal of this study is to develop a statistical model to predict NBA All-Star teams selections from regular season statistics. By applying Principal Component Analysis (PCA) for dimension reduction and logistic regression in R, we analyse significant performance metrics such as points (PTS), assists (AST), rebounds (REB), plus/minus (+/-), and advanced statistics such as PER to find their effects on All-Star selection.**

The strategy is to gather statistics of NBA players, data preprocessing, and logistic regression application for prediction of players as All-Star (1) or Non-All-Star (0). The model will be assessed on the basis of different measures such as the confusion matrix, AUC-ROC curve, and precision-recall scores.

Through statistical modeling, the research provides useful information on selection process and predictive accuracy validation of performance-based selection. Its findings can influence player development, team strategy, and sports analytics application in professional basketball.

## Objectives

The objective of this project is to develop a data-driven model that is able to predict NBA All-Star selections based on the statistics of players during the regular season. The main goals are:

- Identifying the most crucial statistical determinants of All-Star selection.
- The application of Principal Component Analysis (PCA) to reduce dimensionality and increase model efficiency.
- Building a logistic regression model to classify players as probable All-Stars or not All-Stars.
- Model accuracy testing through performance metrics like the confusion matrix, AUC-ROC curve, and precision-recall analysis.
- This study aims to enlighten us as to the role of statistical performance in All-Star selection and evaluate whether a model can accurately predict picks based only on the performances of players.

## Description of dataset

The dataset have 1717 observations and 31 variables contains all the NBA players personal stats for season 2021-22 to 2023-24 .

The interpretation of variables are given below

1. **Player** – Name of the NBA player.
2. **All\_star** – Indicator (1 = Selected, 0 = Not Selected) for whether the player made the All-Star team.
3. **Team** – The team the player played for in that season.
4. **Season** – The NBA season (e.g., 2023-24).
5. **Age** – Player's age during the season.
6. **GP (Games Played)** – Number of games played in the season.
7. **W (Wins)** – Number of games won by the player's team.
8. **L (Losses)** – Number of games lost by the player's team.
9. **Min (Minutes Played)** – Total minutes played in the season.
10. **PTS (Points)** – Total points scored.
11. **FGM (Field Goals Made)** – Number of field goals made.
12. **FGA (Field Goals Attempted)** – Number of field goals attempted.
13. **FG% (Field Goal Percentage)** – Percentage of successful field goals.
14. **3PM (Three-Point Field Goals Made)** – Number of three-pointers made.
15. **3PA (Three-Point Field Goals Attempted)** – Number of three-pointers attempted.
16. **3P% (Three-Point Percentage)** – Percentage of successful three-pointers.
17. **FTM (Free Throws Made)** – Number of free throws made.
18. **FTA (Free Throws Attempted)** – Number of free throws attempted.
19. **FT% (Free Throw Percentage)** – Percentage of successful free throws.
20. **OREB (Offensive Rebounds)** – Number of offensive rebounds.
21. **DREB (Defensive Rebounds)** – Number of defensive rebounds.
22. **REB (Total Rebounds)** – Total rebounds (OREB + DREB).
23. **AST (Assists)** – Number of assists made.
24. **TOV (Turnovers)** – Number of times the player lost possession.
25. **STL (Steals)** – Number of times the player stole the ball.
26. **BLK (Blocks)** – Number of shots blocked.
27. **PF (Personal Fouls)** – Number of personal fouls committed.
28. **FP (Fantasy Points)** – Fantasy basketball points based on player stats.
29. **DD2 (Double-Doubles)** – Number of games with double figures in two statistical categories.
30. **TD3 (Triple-Doubles)** – Number of games with double figures in three statistical categories.
31. **Plus\_Minus** – Player's impact on the score while on the court.

## Sources of dataset

The data set has taken from the official site of NBA

1. <https://www.nba.com/stats/players/traditional?PerMode=Totals&sort=PTS&dir=-1&Season=2023-24>
2. <https://www.nba.com/stats/players/traditional?PerMode=Totals&sort=PTS&dir=-1&Season=2022-23>
3. <https://www.nba.com/stats/players/traditional?PerMode=Totals&sort=PTS&dir=-1&Season=2021-22>

- ❖ For predicting 2024-2025 NBA all star team we use the 2024-2025 NBA data till 26<sup>th</sup> March 2025

The data has taken from

<https://www.nba.com/stats/players/traditional?PerMode=Totals&sort=PTS&dir=-1&Season=2024-25>

### Combined Data of 3 seasons in CSV format

[https://drive.google.com/file/d/1PgXgDaZ2jp0RhhFBOK8TnkuwXucUIFnn/view?usp=drive\\_link](https://drive.google.com/file/d/1PgXgDaZ2jp0RhhFBOK8TnkuwXucUIFnn/view?usp=drive_link)

### 2024-2025 Data in CSV format

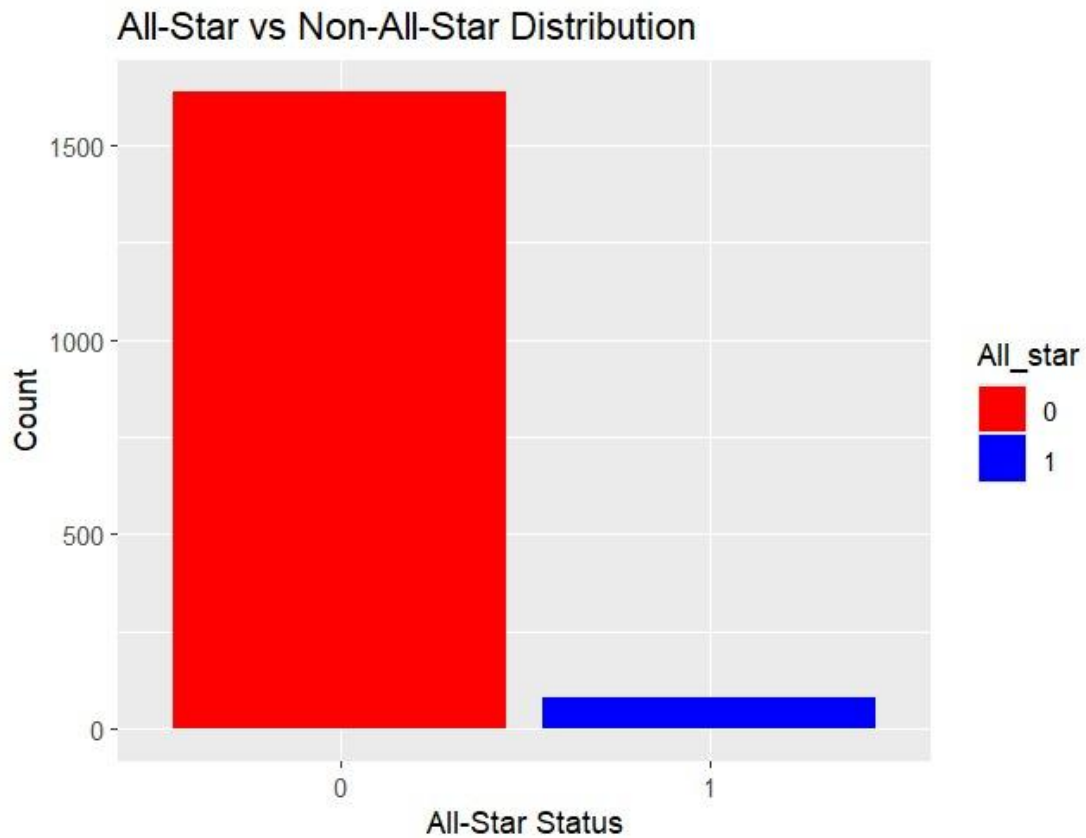
[https://drive.google.com/file/d/1PTmtKapmsPeB-VOTD3IfBKwMNRoUNt57/view?usp=drive\\_link](https://drive.google.com/file/d/1PTmtKapmsPeB-VOTD3IfBKwMNRoUNt57/view?usp=drive_link)

## EDA (Exploratory Data Analysis)

The table below shows the summary of all the variables present in the dataset

| Player           | All.star        |                | Team             | Age             | GP             | W              |
|------------------|-----------------|----------------|------------------|-----------------|----------------|----------------|
| Length:1716      | Min. :0.00000   |                | Length:1716      | Min. :19.00     | Min. : 1.00    | Min. : 0.00    |
| Class :character | 1st Qu.:0.00000 |                | Class :character | 1st Qu.:23.00   | 1st Qu.:23.00  | 1st Qu.: 9.00  |
| Mode :character  | Median :0.00000 |                | Mode :character  | Median :25.00   | Median :51.00  | Median :22.00  |
|                  | Mean :0.04604   |                |                  | Mean :26.16     | Mean :45.65    | Mean :22.91    |
|                  | 3rd Qu.:0.00000 |                |                  | 3rd Qu.:29.00   | 3rd Qu.:68.00  | 3rd Qu.:36.00  |
|                  | Max. :1.00000   |                |                  | Max. :43.00     | Max. :84.00    | Max. :64.00    |
| L                | Min             | PTS            | FGM              | FGA             | FG_PCT         |                |
| Min. : 0.00      | Min. : 0.7      | Min. : 0.0     | Min. : 0.0       | Min. : 0.0      | Min. : 0.00    |                |
| 1st Qu.:11.00    | 1st Qu.: 231.9  | 1st Qu.: 82.0  | 1st Qu.: 31.0    | 1st Qu.: 69.0   | 1st Qu.: 40.38 |                |
| Median :24.00    | Median : 912.0  | Median : 328.0 | Median :123.0    | Median : 266.5  | Median : 44.90 |                |
| Mean :22.74      | Mean :1038.7    | Mean : 486.7   | Mean :178.9      | Mean : 380.3    | Mean : 44.77   |                |
| 3rd Qu.:32.00    | 3rd Qu.:1769.6  | 3rd Qu.: 749.0 | 3rd Qu.:276.0    | 3rd Qu.: 579.0  | 3rd Qu.: 50.00 |                |
| Max. :65.00      | Max. :2988.6    | Max. :2370.0   | Max. :837.0      | Max. :1652.0    | Max. :100.00   |                |
| X3PM             | X3PA            | X3P_PCT        | FTM              | FTA             | FT_PCT         |                |
| Min. : 0.00      | Min. : 0.0      | Min. : 0.00    | Min. : 0.00      | Min. : 0.00     | Min. : 0.00    |                |
| 1st Qu.: 3.00    | 1st Qu.: 14.0   | 1st Qu.: 25.50 | 1st Qu.: 9.00    | 1st Qu.: 13.00  | 1st Qu.: 66.30 |                |
| Median : 30.00   | Median : 91.0   | Median : 33.30 | Median : 40.00   | Median : 54.00  | Median : 75.90 |                |
| Mean : 53.93     | Mean :149.8     | Mean : 29.85   | Mean : 75.06     | Mean : 96.21    | Mean : 69.78   |                |
| 3rd Qu.: 86.00   | 3rd Qu.:242.0   | 3rd Qu.: 37.90 | 3rd Qu.:100.00   | 3rd Qu.:132.25  | 3rd Qu.: 83.40 |                |
| Max. :357.00     | Max. :876.0     | Max. :100.00   | Max. :669.00     | Max. :803.00    | Max. :100.00   |                |
| OREB             | DREB            | REB            | AST              | TOV             | STL            | BLK            |
| Min. : 0.0       | Min. : 0.0      | Min. : 0.0     | Min. : 0.0       | Min. : 0.00     | Min. : 0.0     | Min. : 0.00    |
| 1st Qu.: 9.0     | 1st Qu.: 30.0   | 1st Qu.: 40.0  | 1st Qu.: 17.0    | 1st Qu.: 10.00  | 1st Qu.: 7.0   | 1st Qu.: 3.00  |
| Median : 29.0    | Median :109.0   | Median : 143.5 | Median : 65.0    | Median : 40.00  | Median : 26.0  | Median : 12.00 |
| Mean : 44.9      | Mean :143.5     | Mean : 188.4   | Mean :109.9      | Mean : 56.46    | Mean : 32.1    | Mean : 20.80   |
| 3rd Qu.: 61.0    | 3rd Qu.:215.2   | 3rd Qu.: 279.2 | 3rd Qu.:153.2    | 3rd Qu.: 84.00  | 3rd Qu.: 51.0  | 3rd Qu.: 28.25 |
| Max. :349.0      | Max. :826.0     | Max. :1120.0   | Max. :752.0      | Max. :303.00    | Max. :150.0    | Max. :254.00   |
| PF               | FP              | DD2            | TD3              | X...            |                |                |
| Min. : 0.00      | Min. : -1.0     | Min. : 0.000   | Min. : 0.0000    | Min. : -642.00  |                |                |
| 1st Qu.: 22.75   | 1st Qu.: 188.8  | 1st Qu.: 0.000 | 1st Qu.: 0.0000  | 1st Qu.: -69.00 |                |                |
| Median : 75.00   | Median : 759.0  | Median : 0.000 | Median : 0.0000  | Median : -7.00  |                |                |
| Mean : 83.64     | Mean : 979.9    | Mean : 3.822   | Mean : 0.2249    | Mean : 0.00     |                |                |
| 3rd Qu.:132.00   | 3rd Qu.:1565.5  | 3rd Qu.: 3.000 | 3rd Qu.: 0.0000  | 3rd Qu.: 54.25  |                |                |
| Max. :286.00     | Max. :4609.0    | Max. :77.000   | Max. :29.0000    | Max. : 682.00   |                |                |





|          |              |
|----------|--------------|
| All Star | Not All star |
| 79       | 1637         |

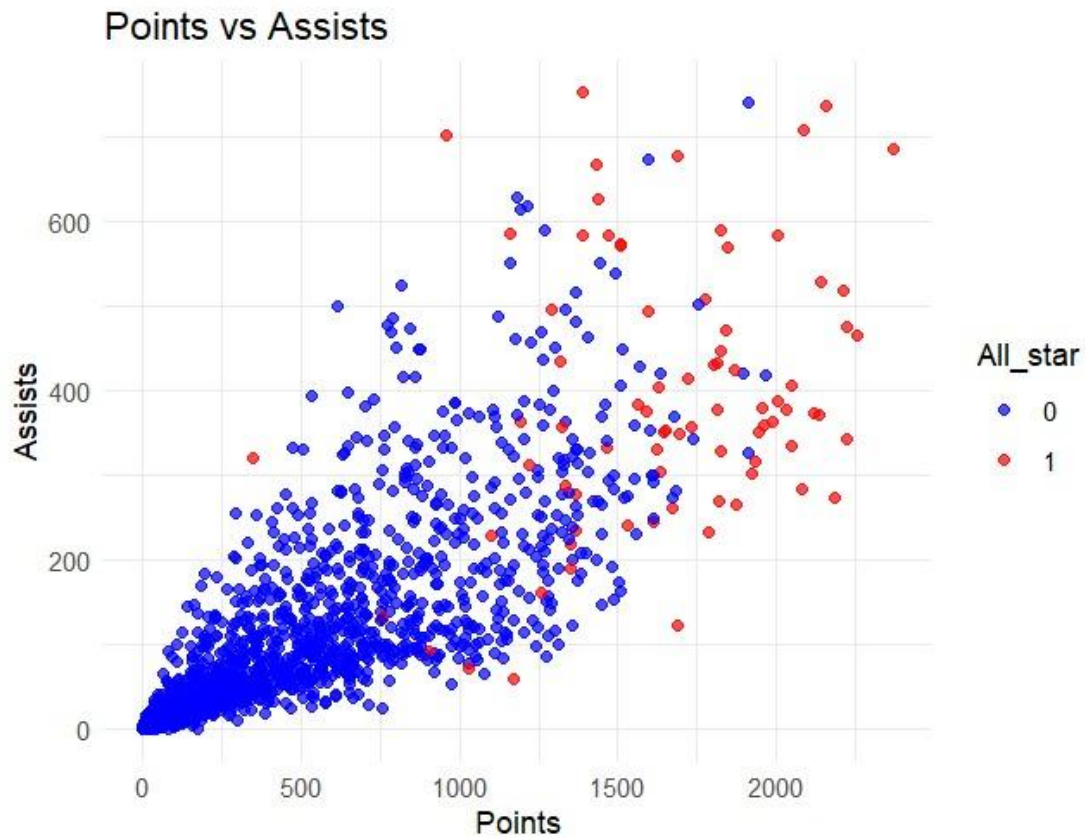
**Distribution of All-Star vs. Non-All-Star Players:**

- A bar chart shows the number of players classified as All-Stars versus Non-All-Stars, helping to understand the class imbalance.
- From the given data, there are **79 All-Stars and 1637 Non-All-Stars**, indicating a highly imbalanced dataset.

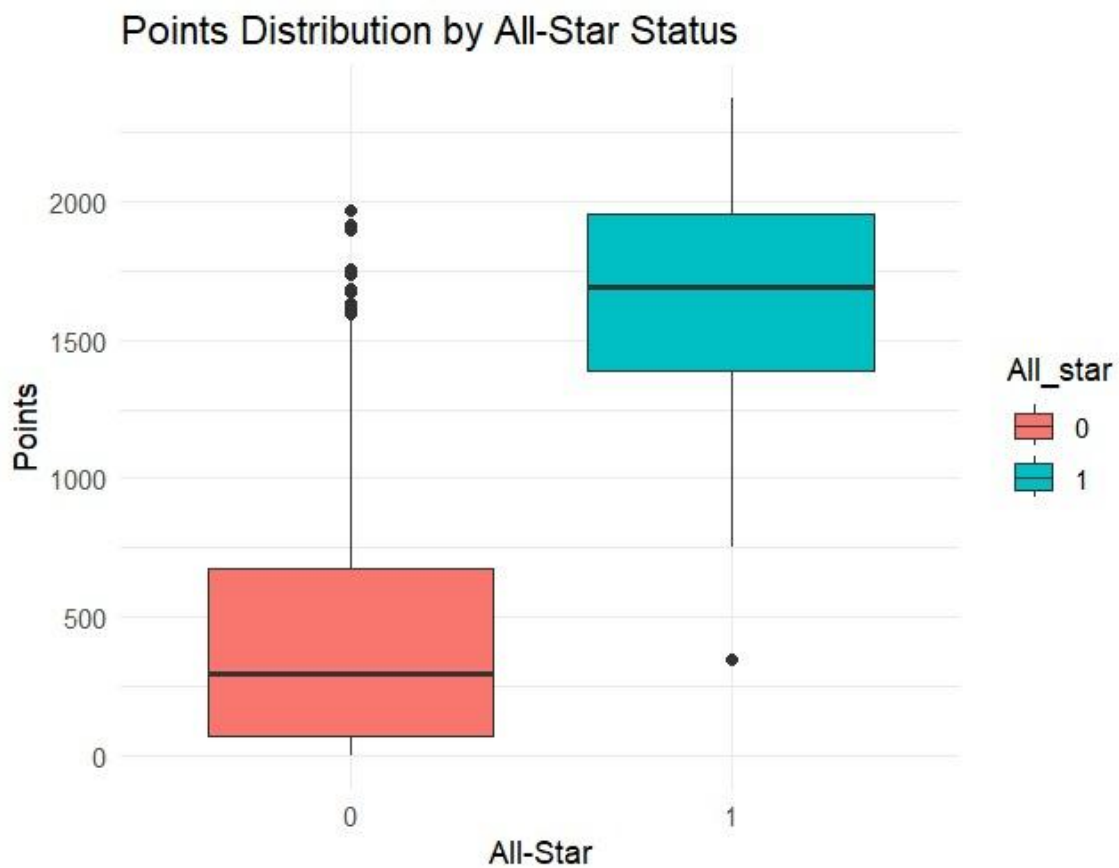
| Player                  | Season  | PTS  | All_star |
|-------------------------|---------|------|----------|
| Luka Doncic             | 2023-24 | 2370 | 1        |
| Shai Gilgeous-Alexander | 2023-24 | 2254 | 1        |
| Jayson Tatum            | 2022-23 | 2225 | 1        |
| Giannis Antetokounmpo   | 2023-24 | 2222 | 1        |
| Jalen Brunson           | 2023-24 | 2212 | 1        |
| Joel Embiid             | 2022-23 | 2183 | 1        |
| Trae Young              | 2021-22 | 2155 | 1        |
| Luka Doncic             | 2022-23 | 2138 | 1        |
| Shai Gilgeous-Alexander | 2022-23 | 2135 | 1        |
| DeMar DeRozan           | 2021-22 | 2118 | 1        |
| Nikola Jokic            | 2023-24 | 2085 | 1        |
| Joel Embiid             | 2021-22 | 2079 | 1        |
| Anthony Edwards         | 2023-24 | 2049 | 1        |
| Jayson Tatum            | 2021-22 | 2046 | 1        |
| Kevin Durant            | 2023-24 | 2032 | 1        |
| Nikola Jokic            | 2021-22 | 2004 | 1        |
| Giannis Antetokounmpo   | 2021-22 | 2002 | 1        |
| Jayson Tatum            | 2023-24 | 1987 | 1        |
| De'Aaron Fox            | 2023-24 | 1966 | 0        |
| Giannis Antetokounmpo   | 2022-23 | 1959 | 1        |
| Stephen Curry           | 2023-24 | 1956 | 1        |
| Anthony Edwards         | 2022-23 | 1946 | 1        |
| Julius Randle           | 2022-23 | 1936 | 1        |
| Donovan Mitchell        | 2022-23 | 1922 | 1        |
| Trae Young              | 2022-23 | 1914 | 0        |
| Zach LaVine             | 2022-23 | 1913 | 0        |
| DeMar DeRozan           | 2023-24 | 1897 | 0        |
| Anthony Davis           | 2023-24 | 1876 | 1        |
| Damian Lillard          | 2022-23 | 1866 | 1        |
| Luka Doncic             | 2021-22 | 1847 | 1        |

The above list gives the top 30 scorers of 3 seasons .

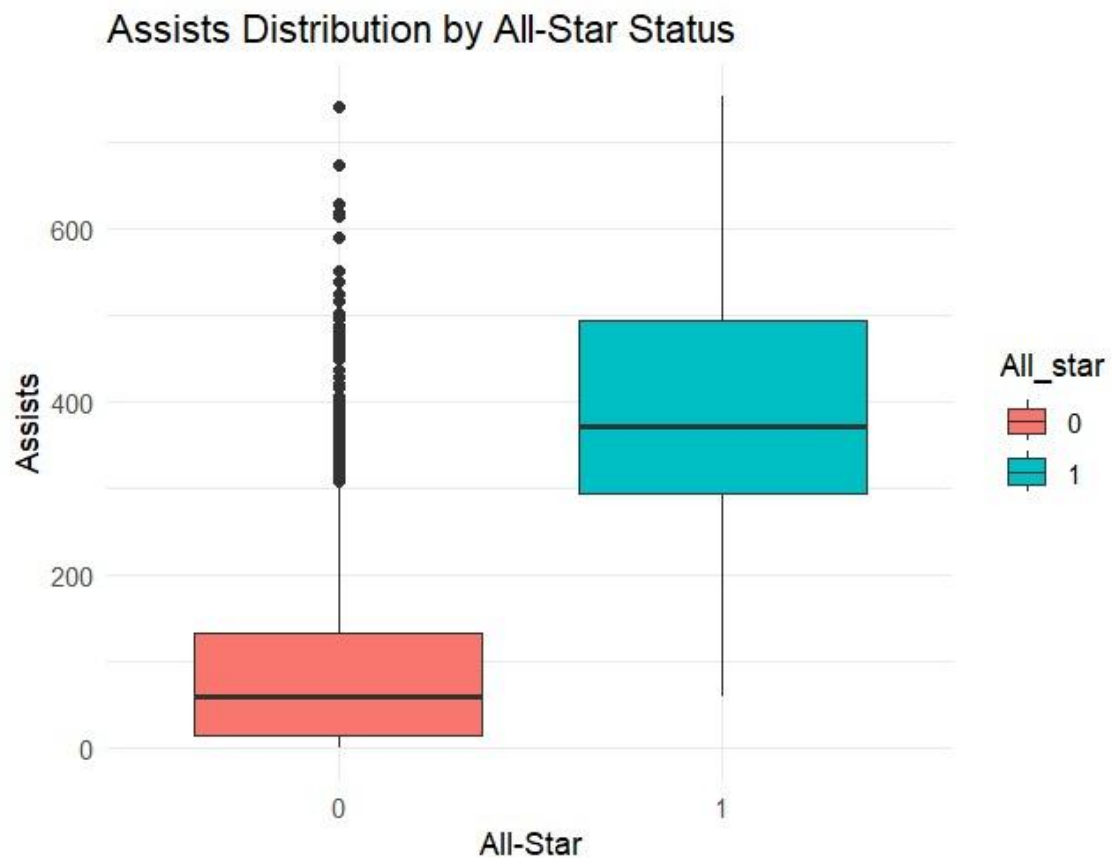
26 players from top 30 scorers list have been selected in All star team this indicates that being in top points scorers list implies a higher probability of being selected for all star team .



The above plot visualizes the relationship between points scored and assists for NBA All star selection. There appears to be a positive correlation between the two variables. As the points scored increases, the variability in the assists also increases. The data indicates that higher assists and points are less rejected .



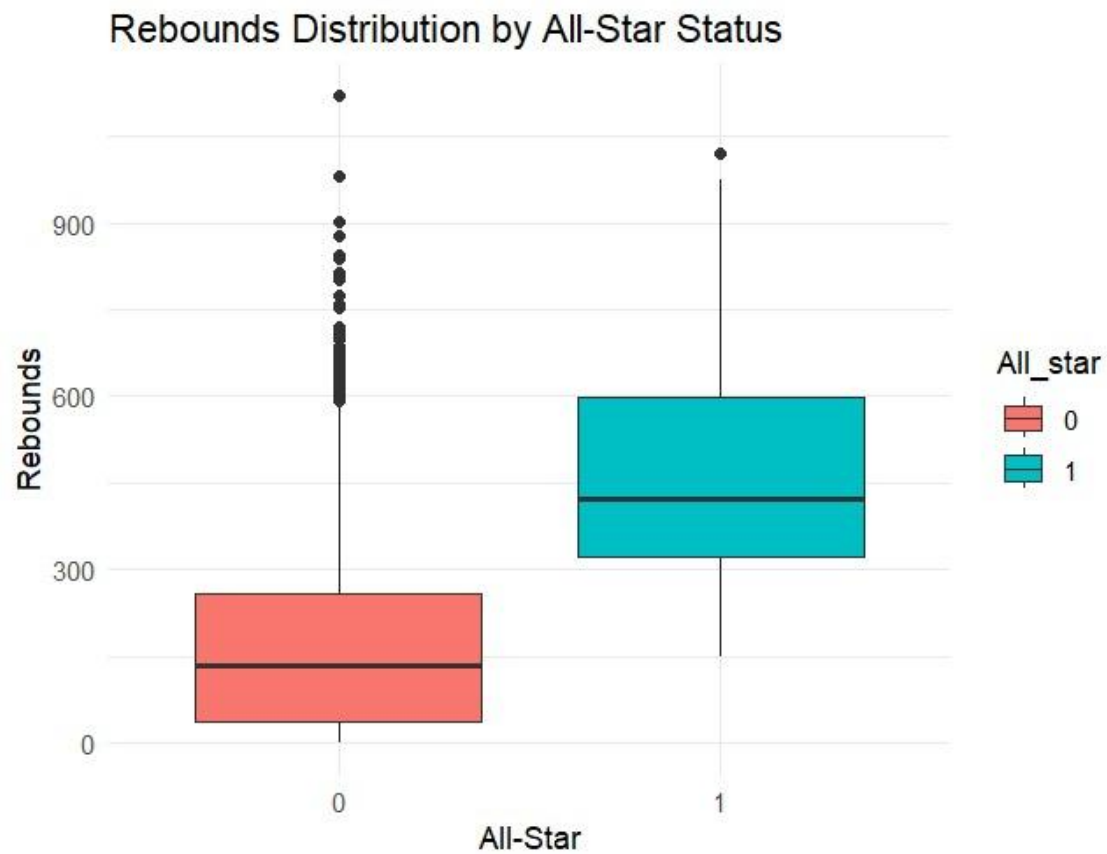
The boxplot shows that All-Star players generally have a significantly higher median and greater spread in points compared to non-All-Stars. Non-All-Star players exhibit more variability, with some outliers scoring high but not making the All-Star team. This suggests that points is a strong predictor of All-Star selection.



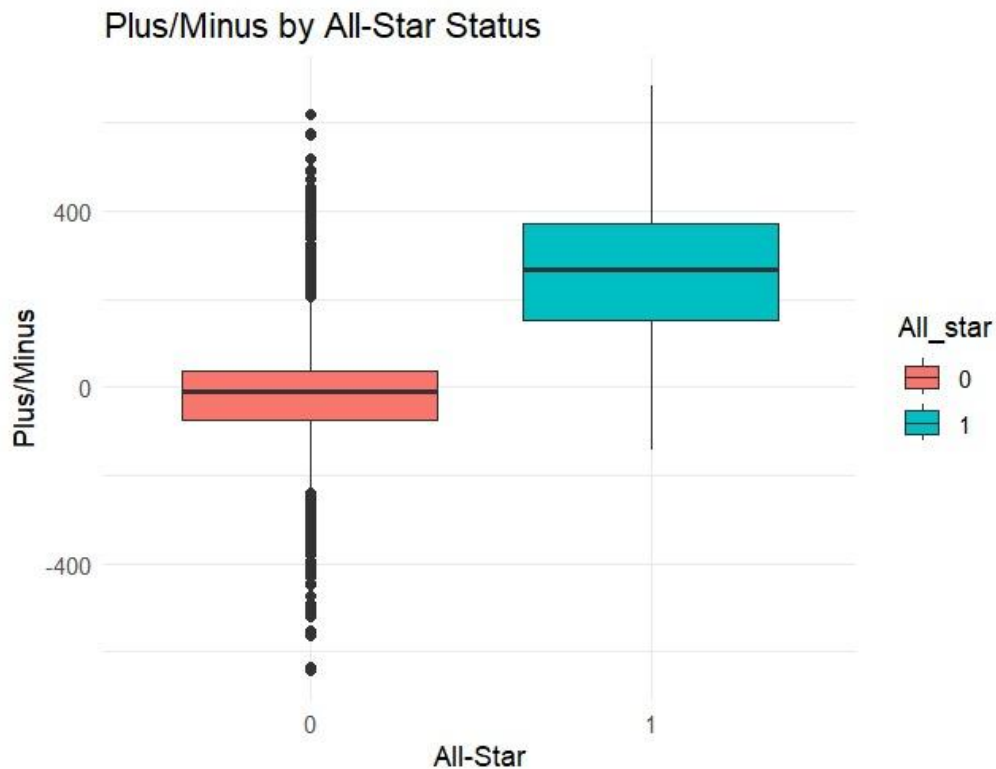
This boxplot compares assists between NBA All stars and non All stars . All-Stars generally record more assists, but the difference is not as pronounced as in points .

While passing ability is valued, it might not be the strongest predictor for selection.

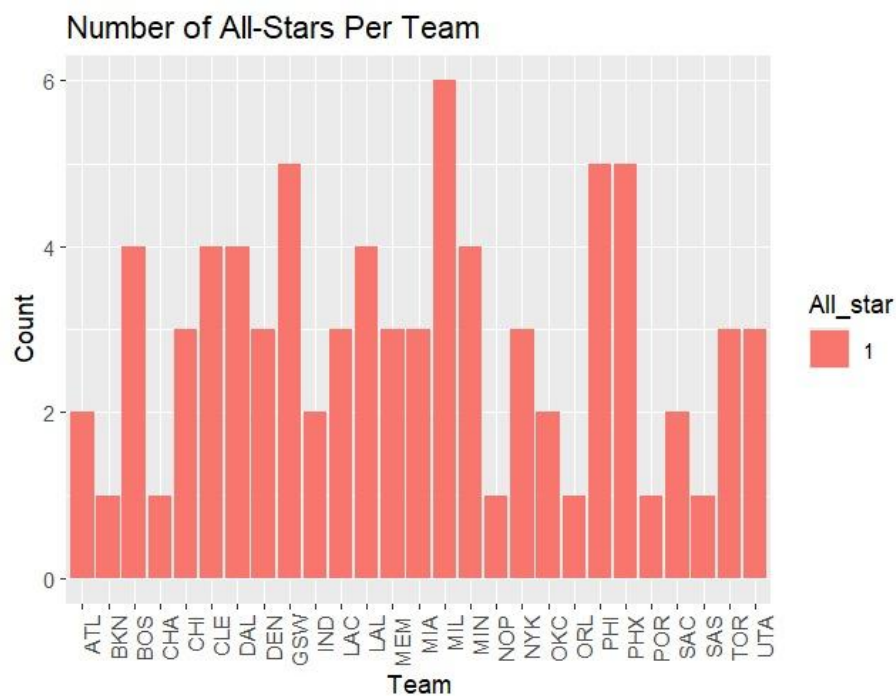
Some Non-All-Stars have high assist numbers, possibly point guards who didn't make the All-Star team.



This boxplot compares rebounds between NBA All-Stars (1) and non-All-Stars (0). All-Stars generally have higher rebounds, with a greater median and a wider distribution. Non-All-Stars have more outliers, indicating some high-rebounding players who were not selected.



This Boxplot compares Plus/Minus effect between All star(1) and non all stars(0) . All-Stars generally have higher Plus/Minus , with a greater median and a wider distribution. Non-All-Stars have more outliers, indicating some players with high Plus/Minus who were not selected.



## Principal Component Analysis

When we are facing a large set of correlated variable principal component allow us to summarize the set with a less number of representative variables that jointly explain most of the variability of the original set . To execute principal component regression we use principal components as predictors in a regression model instead of the wider set of variables .

Principal Component Analysis suggests the technique through which principal components are calculated, and after application of these components in interpreting the data.

PCA is an unsupervised approach, since it includes only a set of Characteristics  $X_1, X_2, \dots, X_p$  and no associated response Y. Beyond providing derived variables to feed supervised learning models, PCA is also a useful data visualization tool to express observations and variables. Moreover, PCA can also be used to perform data imputation to provide values to complete missing entries within a data matrix.

Suppose that we wish to visualize n observations with measurements on a set of p variables . . PCA discovers a low-dimensional representation of a data set that is made up of as much as possible of the variation. The idea is that each of the n observations exists in p-dimensional space, but not all of those dimensions are equally interesting. PCA seeks fewer dimensions which are as interactive as possible, with the definition of interactive being tested by how much the observations differ along each dimension. Each of the dimensions discovered by PCA is a linear combination of the p variables.

The *first principal component* of a set of variables  $X_1, X_2, \dots, X_p$  is the normalized linear combination of the variables.

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

That has largest variation. By normalized we indicate that  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .

We indicate to the elements  $\phi_{11}, \dots, \phi_{p1}$  as the *loadings* of the first principal component; together, the loadings make up the principal component Load-loading vector,

$\phi_1 = (\phi_{11} \ \phi_{21} \ \dots \ \phi_{p1})^T$ . We restrict the loadings so that the sum of their squares equals one. Otherwise, assigning excessively large absolute values to these elements could lead to an arbitrarily large variance.

### Computation of first principal component:

As we are only interested in variance, we assume that each of the variables in X has been centered to have mean zero. We then look for the linear combination of the sample feature values of the form

$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip}$  that has largest sample variance, subject to the constraint that  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .



In other words, the loadings vector for the first principal component is determined by solving the optimization problem.

$$\begin{aligned} & \underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \\ & \text{subject to} \quad \sum_{j=1}^p \phi_{j1}^2 = 1. \end{aligned}$$

### Computation of second principal component :

The second principal component scores  $z_{12}, z_{22}, \dots, z_{n2}$  take the form

$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$  where  $\phi_2$  The second principal component loading vector consists of elements  $\phi_{12}, \phi_{22}, \dots, \phi_{p2}$ . It is found to be that constraining  $Z_2$  to be uncorrelated with

$Z_1$  is equivalent to constraining the direction  $\phi_2$  to be orthogonal to the direction  $\phi_1$ .

### Generalization of first M principal components :

Similarly the first M principal component score vectors and the first M principal component loading vectors deliver the best M-dimensional estimation (in terms of Euclidean distance) to the  $i^{\text{th}}$  observation  $x_{ij}$ . This representation can be written as

$$x_{ij} \approx \sum_{m=1}^M z_{im} \phi_{jm}.$$

This means that the smallest possible value of the above equation is

$$\sum_{j=1}^p \sum_{i=1}^n \left( x_{ij} - \sum_{m=1}^M z_{im} \phi_{jm} \right)^2.$$

In summary, the M principal component score vectors and M principal component loading vectors collectively provide a reliable approximation of the data when M is sufficiently large.

When  $M = \min(n - 1, p)$ , then the

representation is exact:  $x_{ij} = \sum_{m=1}^M z_{im} \phi_{jm}$ .

### Insights from PCA on NBA Data:

Here we have 27 predictor variables  $X_1 = \text{Age}$ ,  $X_2 = \text{GP}$ ,  $X_3 = \text{W}$ ,  $X_4 = \text{L}$ ,  $X_5 = \text{Min}$ ,  $X_6 = \text{PTS}$ ,  $X_7 = \text{FGM}$ ,  $X_8 = \text{FGA}$ ,  $X_9 = \text{FG}\%$ ,  $X_{10} = \text{X3PM}$ ,  $X_{11} = \text{X3PA}$ ,  $X_{12} = \text{X3P}\%$ ,  $X_{13} = \text{FTM}$ ,  $X_{14} = \text{FTA}$ ,  $X_{15} = \text{FT}\%$ ,  $X_{16} = \text{OREB}$ ,  $X_{17} = \text{DREB}$ ,  $X_{18} = \text{REB}$ ,  $X_{19} = \text{AST}$ ,  $X_{20} = \text{TOV}$ ,  $X_{21} = \text{STL}$ ,  $X_{22} = \text{BLK}$ ,  $X_{23} = \text{PF}$ ,  $X_{24} = \text{FP}$ ,  $X_{25} = \text{DD2}$ ,  $X_{26} = \text{TD3}$ ,  $X_{27} = \text{Plus/Minus}$ .  
(Variables are explained in description of dataset part)

And 1716 players observations .

The followings are the loading of the corresponding variable .

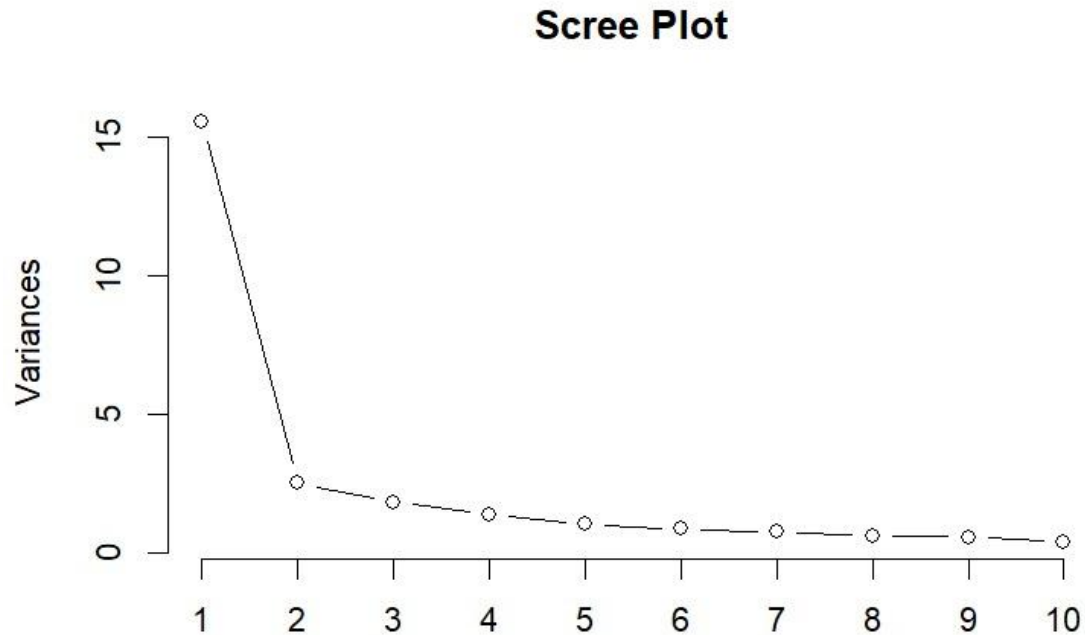
| PC   | Standard Deviation | Proportion of Variance | Cumulative Proportion |
|------|--------------------|------------------------|-----------------------|
| PC1  | 3.9421             | 0.5756                 | 0.5756                |
| PC2  | 1.5885             | 0.09346                | 0.66903               |
| PC3  | 1.35629            | 0.06813                | 0.73716               |
| PC4  | 1.17871            | 0.05146                | 0.78862               |
| PC5  | 1.01918            | 0.03847                | 0.82709               |
| PC6  | 0.93125            | 0.03212                | 0.85921               |
| PC7  | 0.87023            | 0.02805                | 0.88726               |
| PC8  | 0.78787            | 0.02299                | 0.91025               |
| PC9  | 0.76528            | 0.02169                | 0.93194               |
| PC10 | 0.63998            | 0.01517                | 0.94711               |
| PC11 | 0.56661            | 0.01189                | 0.959                 |
| PC12 | 0.54103            | 0.01084                | 0.96984               |
| PC13 | 0.44193            | 0.00723                | 0.97707               |
| PC14 | 0.37558            | 0.00522                | 0.9823                |
| PC15 | 0.36421            | 0.00491                | 0.98721               |
| PC16 | 0.32164            | 0.00383                | 0.99104               |
| PC17 | 0.29335            | 0.00319                | 0.99423               |
| PC18 | 0.25882            | 0.00248                | 0.99671               |
| PC19 | 0.22843            | 0.00193                | 0.99864               |
| PC20 | 0.1468             | 0.00154                | 0.9994                |
| PC21 | 0.09165            | 0.00031                | 0.99975               |
| PC22 | 0.06948            | 0.00018                | 0.99993               |
| PC23 | 0.04313            | 0.00008                | 1                     |
| PC24 | 0.0002876          | 0                      | 1                     |
| PC25 | 0.0002008          | 0                      | 1                     |
| PC26 | 9.47E-16           | 0                      | 1                     |
| PC27 | 8.15E-16           | 0                      | 1                     |

A common threshold for principal component selection is 85-90% cumulative variance. Based on this criterion, retaining 6 is appropriate, as they explain approximately 85.92% of the total variance in the dataset.

**Thus, for further analysis and logistic regression modeling, we select the first 6 principal components.**

### Scree Plot :

In Multivariate statistics a scree plot is a line plot of the eigenvalues of principal components. The scree plot is used to decide the number of principal components to keep in a PCA . Raymond B. Cattell introduces the scree plot in 1966



Scree Plot of NBA DATA

|            | PC1        | PC2          | PC3         | PC4          | PC5         | PC6           |
|------------|------------|--------------|-------------|--------------|-------------|---------------|
| Age        | 0.03525971 | -0.055150609 | 0.06914435  | -0.473535307 | 0.04551341  | -0.8363472201 |
| GP         | 0.21623142 | -0.068862086 | -0.29622927 | -0.061003668 | -0.05103209 | -0.0359977666 |
| W          | 0.19571654 | -0.051938925 | -0.14312424 | -0.352424901 | -0.14519456 | 0.1178098091  |
| L          | 0.17709135 | -0.067895112 | -0.38093783 | 0.278938620  | 0.06776606  | -0.1959102797 |
| Min        | 0.24497726 | -0.075860747 | -0.08126958 | 0.005722380  | -0.09310693 | -0.0266825838 |
| PTS        | 0.24325855 | -0.073520990 | 0.11629263  | 0.065622737  | -0.02906544 | 0.0474679525  |
| FGM        | 0.24420429 | -0.040861115 | 0.09495313  | 0.063010005  | -0.02928517 | 0.0450176478  |
| FGA        | 0.24077353 | -0.119857054 | 0.09272429  | 0.090909525  | -0.04927078 | 0.0292138797  |
| FG_PCT     | 0.07992944 | 0.222579895  | -0.25752279 | -0.223431884 | 0.39427685  | 0.2152553870  |
| X3PM       | 0.18715492 | -0.360104629 | 0.03583344  | -0.016953799 | -0.12057306 | 0.0100400550  |
| X3PA       | 0.19211420 | -0.352271997 | 0.03466362  | 0.021905648  | -0.12260230 | -0.0008733837 |
| X3P_PCT    | 0.07541157 | -0.295118555 | -0.13285053 | -0.174251138 | 0.47753255  | 0.1534244149  |
| FTM        | 0.21719579 | 0.006632835  | 0.21758772  | 0.111907846  | 0.03680685  | 0.0688338738  |
| FTA        | 0.21989466 | 0.051840733  | 0.19835040  | 0.119828720  | 0.03136914  | 0.0626578692  |
| FT_PCT     | 0.10328505 | -0.193952545 | -0.20562985 | -0.121470335 | 0.46705976  | 0.0840741715  |
| OREB       | 0.16632380 | 0.396112157  | -0.19755989 | -0.032651932 | -0.05624551 | -0.0276943410 |
| DREB       | 0.22939174 | 0.204951350  | -0.02077370 | -0.006017199 | -0.01091991 | -0.0552170845 |
| REB        | 0.22097944 | 0.266758051  | -0.07112607 | -0.013719778 | -0.02405519 | -0.0497354983 |
| AST        | 0.20935586 | -0.108412878 | 0.22121639  | 0.049665097  | 0.09650429  | -0.0789636377 |
| TOV        | 0.23357696 | -0.027620519 | 0.14121058  | 0.136192086  | 0.05086415  | -0.0332955547 |
| STL        | 0.21849218 | -0.092063512 | -0.03793020 | -0.028417956 | -0.09690544 | 0.0015259282  |
| BLK        | 0.16171933 | 0.302490115  | -0.18672636 | -0.076998300 | -0.15894490 | 0.0548569888  |
| PF         | 0.22534325 | 0.053002591  | -0.20471391 | -0.007126219 | -0.10414202 | -0.0288559700 |
| FP         | 0.25128380 | 0.021409394  | 0.06429943  | 0.024784627  | -0.02744658 | 0.0038265819  |
| DD2        | 0.16575193 | 0.337255096  | 0.23296979  | 0.062725970  | 0.17277884  | -0.0635186800 |
| TD3        | 0.08639450 | 0.162266453  | 0.38610899  | 0.032299913  | 0.43894870  | -0.0907811625 |
| Plus_Minus | 0.07124119 | 0.032492662  | 0.29933676  | -0.629352302 | -0.19688594 | 0.3544659019  |

### Principal Component Loadings

The principal component loadings indicate the contribution of each original variable to the respective principal components. Higher absolute values represent stronger influences on that component.

## Logistic regression

Logistic regression is a widely used statistical method for binary classification problems. In this study, we will use logistic regression to predict whether an NBA player will be selected for an All-Star Team based on their regular-season performance statistics. Given the binary nature of our target variable (All-Star team selection: Yes or No), logistic regression is an appropriate choice.

The response variable All-Star team selection is denoted by  $Y$  and  $U_1, U_2, U_3, U_4, U_5$  and  $U_6$  are the explanatory variables.

$Y$  represents whether the player will select in a NBA all-star team or not as 1 or 0.

$U_1$  : First principal component

$U_2$  : Second principal component

$U_3$  : Third principal component

$U_4$  : Forth principal component

$U_5$  : Fifth Principal component

$U_6$  : Sixth Principal Component

Suppose the linear relationship is defined as

$$Y = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5 + \beta_6 U_6 + \epsilon$$

The term  $\epsilon$  denotes the error term.

In general,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i} + \beta_5 U_{5i} + \beta_6 U_{6i} + \epsilon_i, \quad i = 1(1)n \\ &= U_i \beta + \epsilon_i \end{aligned}$$

Here,  $y$  can take only 2 values, 0 or 1.

Therefore, it can be assumed that  $y_i \sim \text{Bernoulli}(\pi_i)$

So,  $Y_i$  can take values 1 with probability  $\pi_i$  or 0 with probability  $1 - \pi_i \quad \forall i = 1(1)n$

i.e.  $P(Y_i = 1 | U_1, U_2, U_3, U_4, U_5, U_6) = \pi_i$

As,  $0 < \pi_i < 1$ , we have  $0 < E[Y_i] < 1$

Also,  $\epsilon_i = 1 - U_i \beta$ , whenever  $y_i$  takes 1

and  $\epsilon_i = -U_i \beta$ , whenever  $y_i$  takes 0

$$E[\epsilon_i] = 0$$

In linear regression, the error term is assumed to follow a normal distribution because the response variable is continuous. However, in this case, the response variable is binary (taking values of 0 or 1), making the normality assumption invalid. Instead, we model  $\pi_i$  appropriately based on the predictors, allowing it to vary across different units.

We use the logit link which is given as:  $\log\left(\frac{\pi_i}{1 - \pi_i}\right)$

$$= \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i} + \beta_5 U_{5i} + \beta_6 U_{6i}, \quad \forall i = 1(1)n.$$

We can also write,  $\pi_i$

$$\begin{aligned} &= \frac{\exp(\beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i} + \beta_5 U_{5i} + \beta_6 U_{6i})}{1 + \exp(\beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i} + \beta_5 U_{5i} + \beta_6 U_{6i})} \\ &= h(U_i) \text{ some function} \end{aligned}$$

#### Maximum Likelihood Estimation of Parameters

The probability density function (p.d.f) of  $y$  is given by  $f_i(y_i)$

$$= \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad y_i \text{ takes 1 or 0.}$$

The likelihood function is given by:

$$\begin{aligned} L(\beta) = L &= \prod_{i=1}^n f_i(y_i) \\ &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \end{aligned}$$

Taking log on both sides,

$$\log(L) = \sum_{i=1}^n [\ln(\pi_i) \cdot y_i + (1 - y_i) \cdot \ln(1 - \pi_i)]$$

The parameters

$\beta_0, \beta_1, \beta_2, \dots, \beta_6$  can be estimated by the Fisher scoring method. The score equations are described as follows:

$$\sum_{i=1}^n (y_i - \pi_i) = 0$$

$$\sum_{i=1}^n (y_i - \pi_i) \cdot U_{1i} = 0$$

.....

.....

$$\sum_{i=1}^n (y_i - \pi_i) \cdot U_{6i} = 0$$

These equations need to be solved numerically to obtain estimates of the parameters  $\beta_0, \beta_1, \beta_2, \dots, \beta_6$ .

The fitted values are then obtained as:  $\hat{y}_i = \hat{\pi}_i = \frac{1}{1 + \exp(-\hat{\eta}_i)}$

Coefficients:

|             | Estimate | Std. Error | z value | Pr(> z ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | -6.36025 | 0.57410    | -11.079 | < 2e-16  | *** |
| U1          | 0.48409  | 0.07421    | 6.523   | 6.9e-11  | *** |
| U2          | 0.08896  | 0.08321    | 1.069   | 0.2850   |     |
| U3          | 1.36096  | 0.20169    | 6.748   | 1.5e-11  | *** |
| U4          | -0.22602 | 0.14642    | -1.544  | 0.1227   |     |
| U5          | -0.47188 | 0.26589    | -1.775  | 0.0759   | .   |
| U6          | 0.35538  | 0.20876    | 1.702   | 0.0887   | .   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

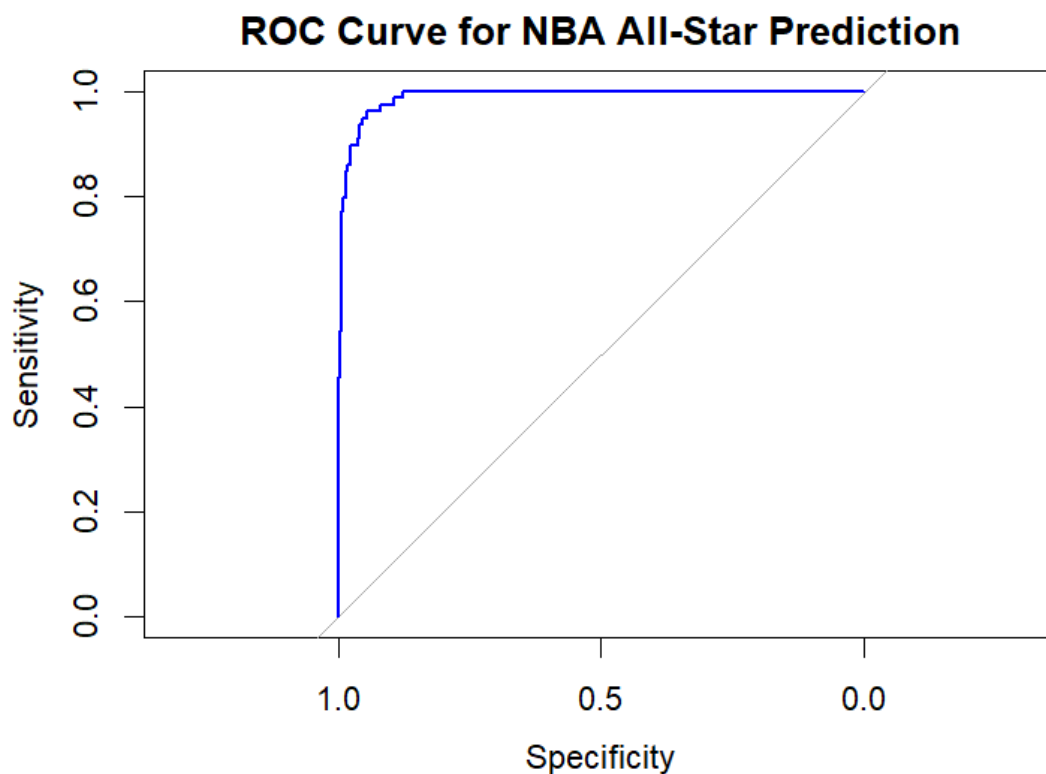
Null deviance: 640.68 on 1715 degrees of freedom  
 Residual deviance: 194.14 on 1709 degrees of freedom  
 AIC: 208.14

Number of Fisher Scoring iterations: 9

The above diagram represents the fitted logistic regression model with logit link with the corresponding standard errors.

### ROC curve and other performance metrics

ROC curve helps in measuring the performance of a model. It plots the True positive Rate(sensitivity) and False positive Rate (Specificity) on y-axis and x-axis for every possible cutoff value. A change in possible threshold value will change both the metrics TPR and FPR. A perfect classifier is one which has a high TPR and low FPR. The diagonal line in the ROC curve represents the chance line where the probability of getting the response as a success when it is a success is as likely as when it is not a success. True Positive rate is the proportion of positive observation which are correctly classified. Ideally one would want a higher FPR. False positive rate is the proportion of negative observation which the model incorrectly identifies as positive. Ideally one would want a lower FPR. Area under the curve (AUC) represents the area under the ROC curve. Higher value of AUC means the model can differentiate appropriately between the positive and negative cases.





**AUC : 0.989135**

The curve above represents the ROC with a high value of AUC. The threshold probability is selected as the value of probability for which the value of  $TPR \cdot (1 - FPR)$  is maximum. 23  
Threshold probability = 0.989135



The corresponding confusion matrix is given as:

| Predicted  |      |    |
|---|------|----|
| Actual     | 0    | 1  |
| 0   | 1627 | 23 |
| 1   | 10   | 56 |

**True Positive (TP): 56**

**False Positive (FP): 10**

**True Negative (TN): 1627**

**False Negative (FN): 23**

- **True Positive Rate (TPR) / Sensitivity / Recall:**

$$TPR = \left( \frac{TP}{TP+FN} \right) = 0.7088$$

- **False Positive Rate (FPR):**

$$FPR = \left( \frac{FP}{FP+TN} \right) = 0.00610$$

In my model, the **True Positive Rate (TPR) is 0.71**, meaning that the model correctly identifies 71% of actual All-Star players. This indicates that the model has a good ability to recognize players who should be selected as All-Stars based on their performance metrics.

The **False Positive Rate (FPR) is 0.006**, which is extremely low. This means that only **0.6%** of non-All-Star players are incorrectly classified as All-Stars. A low FPR is desirable because it shows that the model rarely misclassifies non-All-Star players as All-Stars.

Overall, this suggests that the model is quite effective in distinguishing between All-Stars and non-All-Stars, as it has a high recall (TPR) and a very low misclassification rate (FPR).

**Deviance**

Deviance can also be considered as a measure of model fit. The deviance function is based on comparing the likelihood of the fitted model and that of the null model.

Let  $D$  represent the deviance function,  $L_1$  denotes the log-likelihood of the fitted model on the given data and  $L_0$  be the log-likelihood of the null model.

The deviance function is then defined as  $D = 2 \times (L_1 - L_0)$

A smaller deviance value indicates a better fit of the logistic model as compared to the model with perfect fit.

The null deviance is 640.68 on 1715 degrees of freedom whereas the deviance after fitting the logistic model is 194.14 on 1709 degrees of freedom. The fall in the deviance value shows that the logistic model can explain some of the variability of response variable.

## Comparing Predicted vs Actual 2025 NBA All-Star Selections

We applied the Logistic regression model in 2024-2025 NBA data

| Player                  | All_star_probability | Actual_All_star |
|-------------------------|----------------------|-----------------|
| Shai Gilgeous-Alexander | 0.999520663          | 1               |
| Anthony Edwards         | 0.929800244          | 1               |
| Nikola Jokic            | 0.9999897            | 1               |
| Jayson Tatum            | 0.996162181          | 1               |
| Giannis Antetokounmpo   | 0.994383586          | 1               |
| Cade Cunningham         | 0.990124668          | 1               |
| Devin Booker            | 0.564070621          | 0               |
| Jalen Brunson           | 0.743746281          | 1               |
| Trae Young              | 0.980672278          | 1               |
| Donovan Mitchell        | 0.909910978          | 1               |
| Karl-Anthony Towns      | 0.821177836          | 1               |
| James Harden            | 0.990748984          | 1               |
| LeBron James            | 0.877326151          | 1               |
| Damian Lillard          | 0.644194971          | 1               |
| Darius Garland          | 0.796387332          | 1               |
| Jalen Williams          | 0.649561927          | 1               |
| Alperen Sengun          | 0.916967338          | 1               |
| Jaylen Brown            | 0.534847126          | 1               |
| Evan Mobley             | 0.718069262          | 1               |
| Domantas Sabonis        | 0.67915692           | 0               |
| Luka Doncic             | 0.842724508          | 0               |

This table compares the predicted All-Star probabilities from the logistic regression model with the actual 2024-2025 NBA All-Star selections.

### Observations:

#### 1. High Prediction Accuracy for Star Players:

- Players like Shai Gilgeous-Alexander (0.9995), Nikola Jokić (0.9999), Jayson Tatum (0.9962), and Giannis Antetokounmpo (0.9944) were correctly predicted as All-Stars with near-perfect probabilities.
- This indicates that the model effectively captures top-performing players who are almost certain All-Star selections.

## 2. Borderline Cases and Misclassifications:

- Devin Booker (0.5641) was predicted with a probability slightly above 0.5 but was not selected as an All-Star (Actual = 0).
- Domantas Sabonis (0.6792) and Damian Lillard (0.6442) also had decent probabilities but were not actual All-Stars.
- Jaylen Brown (0.5348) was close to the threshold but was selected as an All-Star (Actual = 1), meaning the model underestimated his chances.

### Note on NBA All Star Games 2025

- From the 2024-2025 season, the NBA is introducing four All-Star teams instead of two, resulting in fewer players per team.
- Three teams consist of regular All-Star players, while the fourth team, *Candace's Rising Stars*, features young players.
- Excluding *Candace's Rising Stars*, 19 out of the 26 players predicted by my model were selected for the NBA All-Star teams.

## Conclusion

This study aimed to predict NBA All-Star selections using regular season performance data through Principal Component Analysis (PCA) and Logistic Regression. By applying PCA, we reduced the dimensionality of the dataset while retaining the most important features influencing All-Star selection. Logistic regression was then used to classify players as All-Stars or Non-All-Stars based on these principal components.

The model demonstrated strong predictive capability, with a **True Positive Rate (TPR) of 0.71** and a **False Positive Rate (FPR) of 0.006**, indicating that it correctly identified a significant proportion of All-Star players while minimizing false classifications. Additionally, the **ROC curve and AUC score** confirmed the model's effectiveness in distinguishing between selected and non-selected players.

An interesting real-world validation of the model was observed in the 2024-2025 NBA season, where **19 out of the 26 players predicted by the model were officially selected in NBA All-Star Teams**. This highlights the practical relevance of performance-based selection criteria.

While the model provides valuable insights into the role of statistical performance in All-Star selection, external factors such as fan votes, media influence, and team success may still impact the final roster. Future research could incorporate additional qualitative variables and explore advanced machine learning models to enhance prediction accuracy.

Overall, this study underscores the power of data-driven decision-making in sports analytics and provides a framework for objective player evaluation.

## R Codes

The link below provides all the R code used to execute the project

<https://drive.google.com/file/d/1dyS4y2-KUZZDPcQjwr6UE17NRPImT9OB/view?usp=sharing>

## References

1. An Introduction to Statistical Learning- Gareth James , Daniela Witten ,Trevor Hastie , Robert Tibshirani
2. Fundamentals of Statistics, Volume 2:A.M. Goon, M.K. Gupta, B. Dasgupta
3. <https://www.nba.com>