NBA ALL STAR SELECTION USING LOGISTIC REGRESSION

SOUMYA KARMAKAR
REG NO-A01-1142-0720-22
SUPERVSIOR - DR. AYAN CHANDRA

.....

WHY SPORTS ANALYTICS!

- Passion for sports
- Interest in applying statistics to real-world performance
- Sports Analytics: A rising global field
- Basketball & Sports Analytics still emerging in India
- Unique and challenging project choice
- Hope to contribute to the growth of data-driven sports in India

ABOUT NBA & NBA ALL STAR GAMES

- NBA (National Basketball Association): Premier men's basketball league in the world
- 30 teams (29 USA, 1 Canada), known for elite talent and global fanbase
- Highly competitive, data-driven, and commercially successful
- NBA All-Star Game: Annual mid-season exhibition game
- Top players selected based on fan votes, media, and coaches
- Mix of skill, popularity, and performance
- Prestigious recognition for players

OBJECTIVES

- To identify key performance metrics influencing NBA All-Star selection
- To reduce dimensionality using Principal Component Analysis (PCA)
- To build a logistic regression model for All-Star prediction
- To evaluate the model using metrics like confusion matrix, TPR, FPR, and AUC
- To assess how well statistical performance alone can explain selection outcomes

DATA DESCRIPTION

SOURCE OF DATASET

1

Official site of NBA
www.nba.com

TIME PERIOD

2

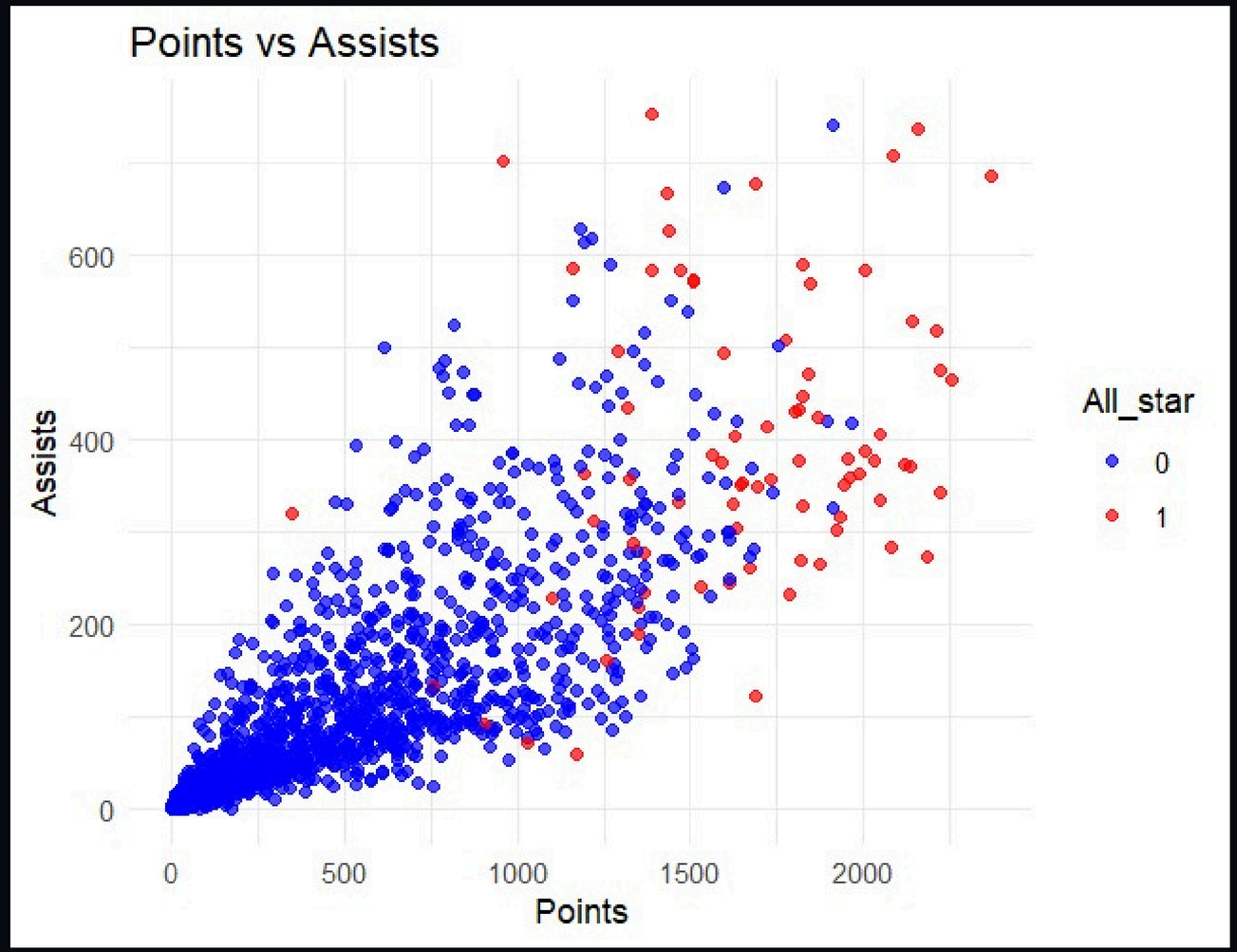
For model building we use 2021-2024 NBA data and for prediction we have use 2024-2025 season data

VARIABLES

3

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.

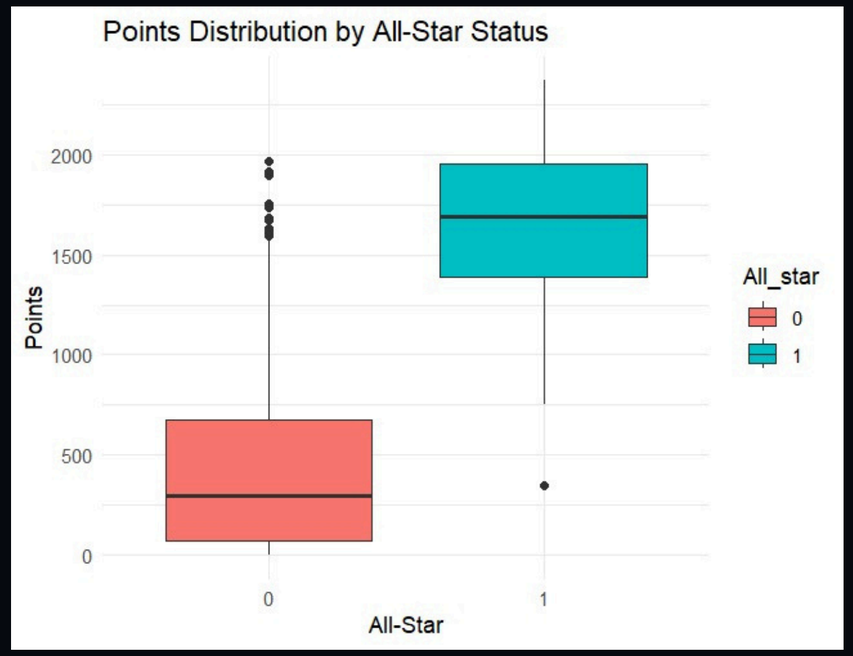
Relationship Between Points and Assists for NBA All-Star Selection

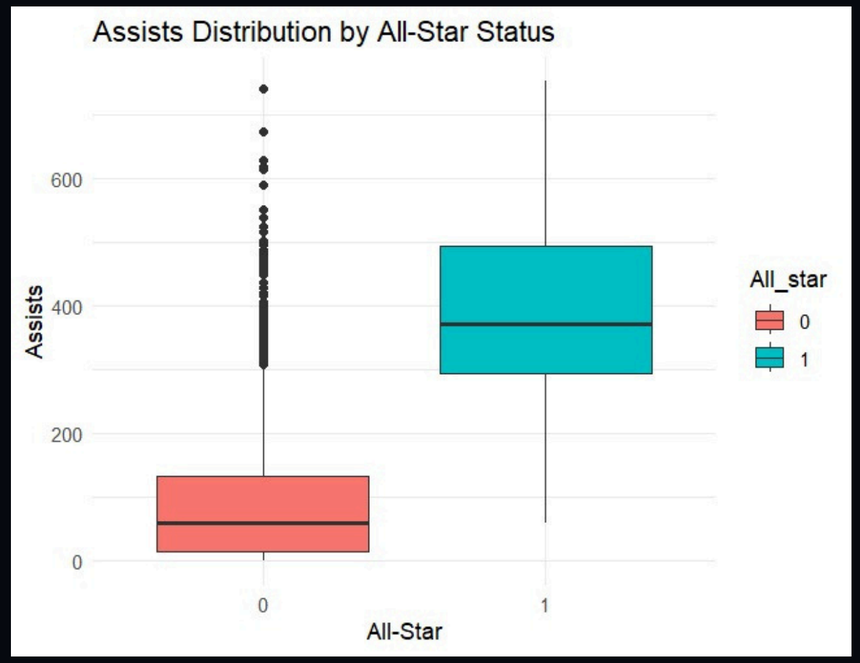


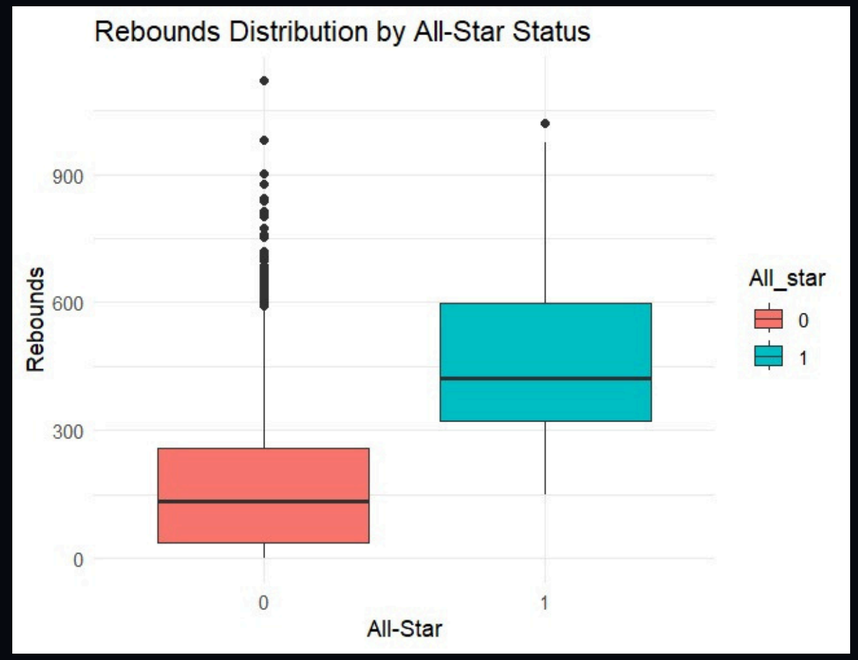
SUMMARY

- There is a positive correlation between points scored and assists.
- As points scored increase, the variability in assists also increases.
- Higher points and assists seem to reduce the likelihood of being rejected for All-Star selection.

EVALUATING KEY PREDICTORS FOR ALL-STAR SELECTION







- **Points:** All-Stars have a higher median and greater spread in points, making it a strong predictor of All-Star selection.
- **Assists:** All-Stars generally record more assists, but the difference is less significant, indicating passing is important but not the strongest predictor.
- **Rebounds:** All-Stars have higher median rebounds with a wider distribution, but non-All-Stars show more outliers, highlighting that rebounds alone don’t guarantee selection.

PCA

(PRINCIPAL COMPONENT ANALYSIS)

WHY PCA ?

- Many performance metrics in NBA data are correlated (e.g., PTS, FGA, FG%)
- Correlation between predictors can lead to multicollinearity, affecting model stability
- PCA reduces dimensionality by converting original variables into uncorrelated components
- Helps in retaining most of the information with fewer variables
- Simplifies the logistic regression model and improves interpretability
- Makes the model more efficient and generalizable

.....

Variance Explained by Principal Components

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

OBSERVATIONS

- PC1 explains 57.56% of total variance.
- Top 6 PCs cumulatively explain ~85.9%.
- Dimensionality reduced from 27 to 6 without major information loss.

- PC1 has the highest variance.
- Sharp drop after PC1, gradual decline from PC2 to PC6.
- Elbow at PC6 → first 6 components are retained

	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

- PCA loadings show how original variables contribute to principal components.
- High loading = strong influence on that component.
- Helps simplify the model without losing important information.

PRINCIPAL COMPONENT LOADINGS

LOGISTIC REGRESSION

LOGISTIC REGRESSION: MODEL EQUATION

The distribution of Y is specified by probabilities of success P(Y=1)= π and the probabilities of failure P(Y=0)=1- π.
E(Y)= π
Then, the multiple logistic regression with binary response is given by:
$Y_i \sim$ Bernoulli(π_i)

$$\pi_i = P(Y_i = 1 \mid U_{1i}, U_{2i}, \ldots, U_{6i}) = \frac{e^{\eta i}}{1+e^{\eta i}} \qquad \text{----------(1)}$$

here, η_i is a function of all the predictor variables,
where $\eta_i = \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \beta_3 U_{3i} + \beta_4 U_{4i} + \beta_5 U_{5i} + \beta_6 U_{6i}$
;i=1(1)1716

REGRESSION TABLE

PARAMETERS	ESTIMATE	STANDARD ERROR	Z VALUE	P VALUE
INTERCEPT	-6.36025	0.57410	-11.079	$<2e^{-16}$
U_1	0.48409	0.07421	6.523	$6.9e^{-11}$
U_2	0.08896	0.08321	1.069	0.2850
U_3	1.36096	0.20169	6.748	$1.5e^{-11}$
U_4	-0.22602	0.14642	-1.544	0.1227
U_5	-0.47188	0.26589	-1.775	0.0759
U_6	0.35538	0.20876	1.702	0.0887

The fitted regression model using logit link is given by -

$$\hat{\pi} = P(Y = 1 \mid U_1, U_2, \ldots, U_6) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}}$$

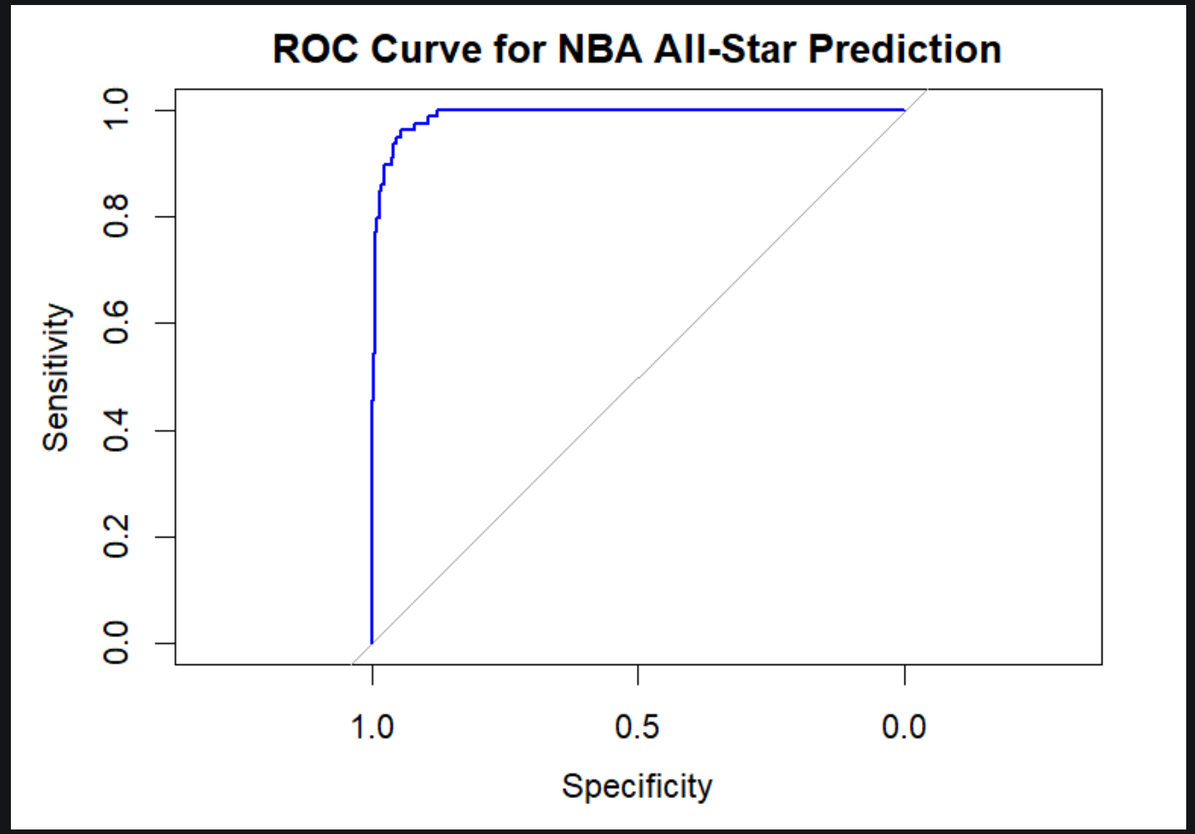
$$\text{where } \hat{\eta} = -6.36025 + 0.48409\,u_1 + 0.08896\,u_2 + 1.36096u_3 - 0.22602u_4 - 0.47188u_5 + 0.35538\,u_6$$

Residual deviance (D)= 194.14

Null deviance (D_0)= 640.68

Since the residual deviance is less than the Null deviance, we can say that the model is fitting well.

RECEIVER OPERATOR CHARACTERISTIC CURVE



- ROC Curve plots True Positive Rate (TPR) vs False Positive Rate (FPR) for various cutoff values.
- A perfect classifier has high TPR and low FPR; the diagonal line represents random guessing.
- AUC (Area Under Curve) measures overall model performance — higher AUC = better classification ability.
- Optimal Threshold is chosen where TPR × (1 – FPR) is maximized.
  → Threshold Probability = 0.989135

QUALITY OF REGRESSION MODEL

Actual \ Predicted	0	1
0	1627	23
1	10	56

🔍 Key Metrics:

- True Positive Rate (Recall): 0.71
- False Positive Rate: 0.006

📝 Interpretation:

- High Recall (71%) → Model effectively identifies actual All-Stars.
- Very low FPR (0.6%) → Rarely misclassifies non-All-Stars.
- Indicates strong performance in All-Star classification.

2024–25 Season Predictions: A Reality Check

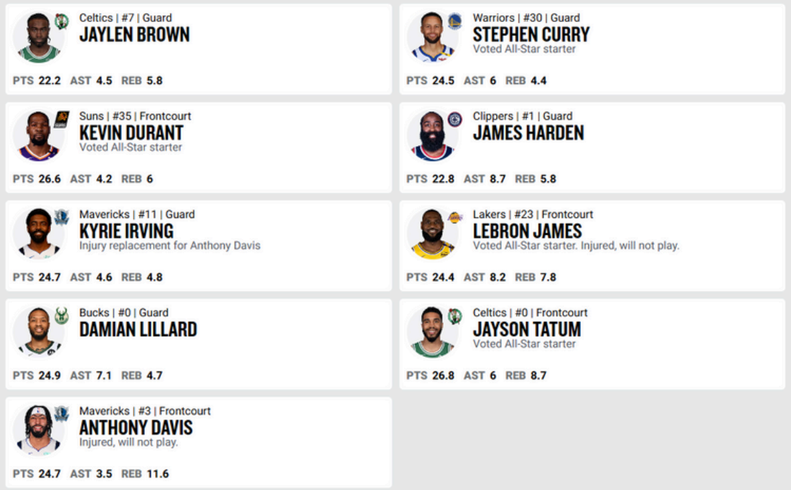
- The model was tested on 2024–2025 NBA player data.
- It predicted 21 players as potential All-Stars based on performance.
- Out of those, 19 players were actually selected as NBA All-Stars.
- This shows the model's predictions matched real-life selections well.
- It proves that performance data alone can be a strong indicator of All-Star selection.

PREDICTED PLAYERS

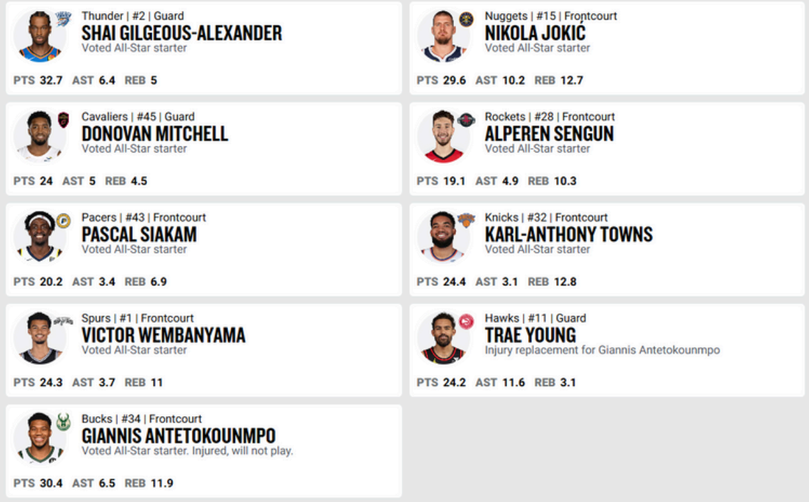
Player	All_star_probability	Actual_All_star
Shai Gilgeous-Alexander	0.999520663	1
Anthony Edwards	0.929800244	1
Nikola Jokic	0.9999897	1
Jayson Tatum	0.996162181	1
Giannis Antetokounmpo	0.994383586	1
Cade Cunningham	0.990124668	1
Devin Booker	0.564070621	0
Jalen Brunson	0.743746281	1
Trae Young	0.980672278	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

ACTUAL TEAMS

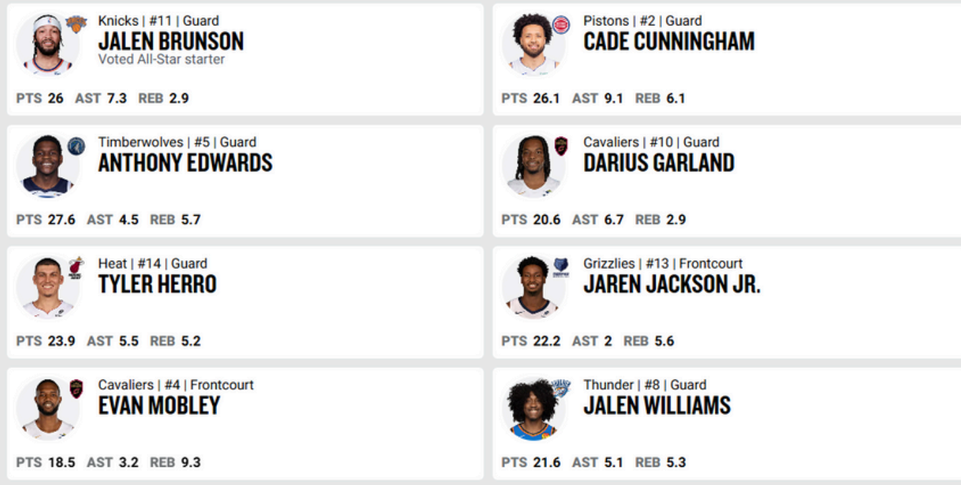
SHAQ'S OGS



CHUCK'S GLOBAL STARS



KENNY'S YOUNG STARS



CONCLUSION

- NBA All-Star selections can be predicted using player performance data.
- PCA helped reduce dimensionality and removed multicollinearity.
- Logistic regression achieved good accuracy with a TPR of 0.70 and FPR of 0.006.
- 19 out of 21 players predicted by the model were actually selected in 2024–25.
- Statistical performance is a strong indicator of All-Star selection.

LIMITATIONS

- All-Star selection is also influenced by fan votes, media attention, and popularity, which are not in the dataset.
- Rookie players and players with fewer games played may be underrepresented.
- Only logistic regression was used — other models like Random Forest or SVM could improve performance.
- The model does not account for position-specific differences.

FUTURE SCOPE

- Integrate non-performance factors like fan votes and media influence.
- Explore advanced models (Random Forest, SVM, XGBoost) to enhance accuracy.
- Use position-based modeling for deeper player role insights.
- Apply the model to early-season data for predicting future All-Stars.
- Extend to other sports and lesser-known leagues, where manual scouting is limited.
- Help selectors re-evaluate overlooked talent using data-driven support.

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

THANK YOU

“QUESTIONS AND FEEDBACK ARE WELCOME!”