Predicting the Next NBA Defensive Player of the Year

Data 144

Austin Ho, Celina Liu, Garima Gupta, Katherine Gan, Momo Siu, Sailesh Kethepelli

TABLE OF CONTENTS

01 BACKGROUND

“Winning is more related to good defense than good offense”

- Jack Ramsay

Teams

	Rk	Team	Overall	Home	Road	E Wins	E Loss	W Wins	W Loss	Conference	Year
0	1	Phoenix Suns	64-18	32-9	32-9	25	5	39	13	West	2021
1	2	Memphis Grizzlies	56-26	30-11	26-15	20	10	36	16	West	2021
2	3	Golden State Warriors	53-29	31-10	22-19	20	10	33	19	West	2021
3	4	Miami Heat	53-29	29-12	24-17	35	17	18	12	East	2021
4	5	Dallas Mavericks	52-30	29-12	23-18	16	14	36	16	West	2021
...	...	...	...	...	...	...	...	...	...	...	...
534	25	Atlanta Hawks	28-54	18-23	10-31	19	35	9	19	East	2004
535	26	Los Angeles Clippers	28-54	18-23	10-31	14	16	14	38	West	2004
536	27	Washington Wizards	25-57	17-24	8-33	16	38	9	19	East	2004
537	28	Chicago Bulls	23-59	14-27	9-32	19	35	4	24	East	2004
538	29	Orlando Magic	21-61	11-30	10-31	17	37	4	24	East	2004

539 rows × 11 columns

Players

	ï»¿	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Year
0	Zylan Cheatham	Zylan Cheatham	SF	26	UTA	1	0	5.0	0.0	3.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2021
1	Zylan Cheatham	Zylan Cheatham	PF	24	NOP	4	0	12.8	1.5	2.3	...	0.8	1.5	2.3	0.8	0.3	0.3	1.0	2.5	3.0	2019
2	Zydrunas Ilgauskas	Zydrunas Ilgauskas	C	35	MIA	72	51	15.9	2.3	4.4	...	1.5	2.5	4.0	0.4	0.3	0.8	0.7	2.6	5.0	2010
3	Zydrunas Ilgauskas	Zydrunas Ilgauskas	C	34	CLE	64	6	20.9	3.0	6.8	...	1.8	3.6	5.4	0.8	0.2	0.8	1.0	2.9	7.4	2009
4	Zydrunas Ilgauskas	Zydrunas Ilgauskas	C	33	CLE	65	65	27.2	5.3	11.1	...	2.4	5.1	7.5	1.0	0.4	1.3	1.4	2.8	12.9	2008
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11138	A.J. Price	A.J. Price	PG	26	WAS	57	22	22.4	2.8	7.2	...	0.4	1.6	2.0	3.6	0.6	0.1	1.1	1.3	7.7	2012
11139	A.J. Price	A.J. Price	PG	25	IND	44	1	12.9	1.3	4.0	...	0.3	1.1	1.4	2.0	0.5	0.0	0.7	0.7	3.9	2011
11140	A.J. Price	A.J. Price	PG	24	IND	50	0	15.9	2.3	6.4	...	0.3	1.1	1.4	2.2	0.6	0.0	1.1	1.2	6.5	2010
11141	A.J. Price	A.J. Price	PG	23	IND	56	2	15.4	2.6	6.3	...	0.2	1.4	1.6	1.9	0.6	0.1	1.1	0.9	7.3	2009
11142	A.J. Hammons	A.J. Hammons	C	24	DAL	22	0	7.4	0.8	1.9	...	0.4	1.3	1.6	0.2	0.0	0.6	0.5	1.0	2.2	2016

11143 rows × 31 columns

DPOY

	Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share	G	MP	...	TRB	AST	STL	BLK	FG%	3P%	FT%	WS	WS/48	Year
1	1	Marcus Smart	27	BOS	37	257	500	0.514	71	32.3	...	3.8	5.9	1.7	0.3	0.418	0.331	0.793	5.6	0.116	2021
2	2	Mikal Bridges	25	PHO	22	202	500	0.404	82	34.8	...	4.2	2.3	1.2	0.4	0.534	0.369	0.834	8.9	0.15	2021
3	3	Rudy Gobert	29	UTA	12	136	500	0.272	66	32.1	...	14.7	1.1	0.7	2.1	0.713	0	0.69	11.7	0.264	2021
4	4	Bam Adebayo	24	MIA	13	128	500	0.256	56	32.6	...	10.1	3.4	1.4	0.8	0.557	0	0.753	7.2	0.188	2021
5	5	Jaren Jackson Jr.	22	MEM	10	99	500	0.198	78	27.3	...	5.8	1.1	0.9	2.3	0.415	0.319	0.823	5.4	0.121	2021
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
371	8T	Shaquille O'Neal	28	LAL	2	2	123	0.016	74	39.5	...	12.7	3.7	0.6	2.8	0.572	0	0.513	14.9	0.245	2000
372	11T	Kobe Bryant	22	LAL	1	1	123	0.008	68	40.9	...	5.9	5	1.7	0.6	0.464	0.305	0.853	11.3	0.196	2000
373	11T	Allen Iverson	25	PHI	1	1	123	0.008	71	42	...	3.8	4.6	2.5	0.3	0.42	0.32	0.814	11.8	0.19	2000
374	11T	Jason Kidd	27	PHO	1	1	123	0.008	77	39.8	...	6.4	9.8	2.2	0.3	0.411	0.297	0.814	9.6	0.15	2000
375	11T	Shawn Marion	22	PHO	1	1	123	0.008	79	36.2	...	10.7	2	1.7	1.4	0.48	0.256	0.81	11.7	0.196	2000

375 rows × 21 columns

WS (Win shares)

	Player	Year	Tm	WS	WS/48
0	A.J. Hammons	2016	DAL	0.0	-0.001
1	A.J. Price	2009	IND	1.2	0.065
2	A.J. Price	2010	IND	0.3	0.020
3	A.J. Price	2011	IND	0.7	0.063
4	A.J. Price	2012	WAS	2.2	0.084
...	...	...	...	...	...
11151	Zydrunas Ilgauskas	2007	CLE	6.1	0.131
11152	Zydrunas Ilgauskas	2009	CLE	2.5	0.088
11153	Zydrunas Ilgauskas	2010	MIA	2.9	0.122
11154	Zylan Cheatham	2019	NOP	0.0	0.034
11155	Zylan Cheatham	2021	UTA	-0.1	-0.610

11156 rows × 5 columns

Important Vocabulary

Win Share

How many wins a player contributes to

True Rebound %

Proportion of available rebounds made by a player

Plus Minus

Box score per 100 possessions against an average player

Block %

Proportion of 2 point field goal attempts blocked by a player

Steal %

Fraction of opponent possessions that end in a steal by a player

VORP

Measure of a player’s contribution above a replacement level player

02 PREPROCESSING

Dataframe Cleaning

PROJECT

DATAFRAMES

DPOY

Took out unnecessary notations (*), Marked who won the award

PLAYERS

Encoded using unicode-escape, renamed column title

TEAMS

Dropped useless columns, found cumulative wins and losses

Additional Preprocessing

WS Dataframe

Found winshare data for players

Acronyms

Updated player's team acronyms based on year

Edge Cases

Fixed edge cases for teams

Redundancies

Removed duplicate entries in player tables

Creating Final Tables

Players

Merged player dataframe with team performance and corresponding winshares

DPOY

Merged the DPOY with the previous player table to serve as predictions we wanted to meet

Filtering

1

Playoffs

Only players on playoff teams are recognized and considered for major awards such as DPOY

Decreased rows in dataframe from 15435 to 5163

2

Games Played

To be recognized for DPOY, a player must play in the majority of games in the season

Filtered another 2841 entries from the dataframe

3

Rebounds Per Game

Rebounds are a major defensive component of the game. We set a filter of 3 rebounds per game

Removed another 677 instances from the dataframe

4

Win Share

The DPOY should contribute positively to wins. We set a filter of 0.5 win shares.

Filtered out just 3 entries from the dataframe

03 MODELING

LINEAR REGRESSION

INSIGHT & ANALYSIS

- Created a model with linear regression to predict the winner: 1 if winner, 0 if not
- Training and Testing Sets were created with Year and Winner as features

MSE:
3.0955488291881096e-30

RSME:
1.7594171845210873e-15

LOGISTIC REGRESSION

KEY FEATURES & INSIGHTS

- For every year, the model was predicting a binary variable: winner
- Training and test datasets were scaled to standardize feature values

Average of all accuracy scores: **0.977**

ACCURACY SCORES

2004	0.923	2013	1.0
2005	1.0	2014	0.923
2006	1.0	2015	1.0
2007	1.0	2016	1.0
2008	1.0	2017	1.0
2009	1.0	2018	1.0
2010	1.0	2019	1.0
2011	0.917	2020	1.0
2012	0.941	2021	0.875

NEURAL NETWORK

DESIGN

Number of features: 51

- Hidden layers: 5
- Solver: lfbgs
- Max iterations: 100
- Learning rate: adaptive
- Initial learning rate = 0.001

MODEL PERFORMANCE

Mean accuracy = 0.932

ACCURACY SCORES

2004	0.692	2013	1.0
2005	1.0	2014	0.923
2006	0.928	2015	0.923
2007	1.0	2016	1.0
2008	0.857	2017	1.0
2009	1.0	2018	1.0
2010	1.0	2019	0.9
2011	0.916	2020	0.833
2012	0.941	2021	0.875

Gradient Boost

Why

- Robust predictive analysis model
- Ensembling method

MSE: 5.396405926331 66e-06

```
from sklearn.ensemble import GradientBoostingRegressor

# Train Test Split and MSE metrics already imported.



parameters = {'n_estimators': 500,
              "max_depth": 4,
              "min_samples_split": 5,
              "learning_rate": 0.01,
              "loss": "absolute_error",
              }
regression = GradientBoostingRegressor(**parameters)
regression.fit(X_train, Y_train)

#metrics
mse = mean_squared_error(Y_test, regression.predict(X_test))
print(mse)
```

Gradient Boost

Visual

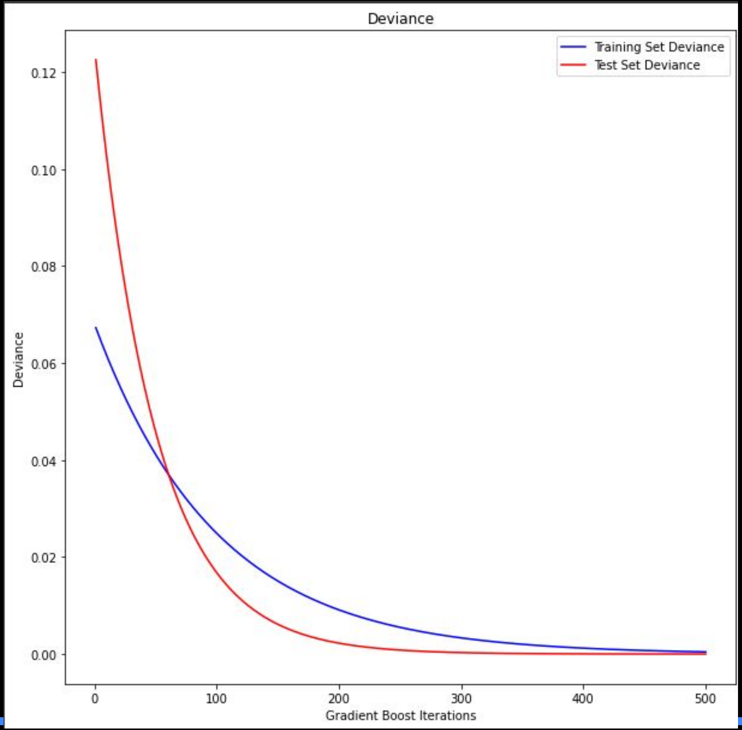
- Optimized within 500 iterations
- Minimized deviance

```python
import matplotlib.pyplot as plt

test_score = np.zeros((parameters["n_estimators"],), dtype=np.float64)
for elem, y_pred in enumerate(regression.staged_predict(X_test)):
  test_score[elem] = mean_squared_error(Y_test, y_pred)

fig = plt.figure(figsize = (10, 10))
plt.subplot(1, 1, 1)
plt.title("Deviance")
plt.plot(np.arange(parameters["n_estimators"]) + 1,
         regression.train_score_,
         "b-",
         label = "Training Set Deviance",
)

plt.plot(np.arange(parameters["n_estimators"]) + 1, test_score, "r-",
         label="Test Set Deviance")

plt.legend(loc = "upper right")
plt.xlabel("Gradient Boost Iterations")
plt.ylabel("Deviance")
plt.show()
```

Gradient Boost

Visual

- Optimized within 500 iterations
- Minimized deviance



04 ANALYSIS

Model Comparison

Logistic Regression

Avg accuracy = 0.977

1 — 2

Neural Network

Avg accuracy = 0.932

Gradient Boost

MSE of 5.396e-06

1 — 2

Linear Regression

MSE of 3.096e-30

Next Steps

Reduced Features

Further analysis on features or more domain knowledge could have allowed us to make more informed decisions on which features to remove

Additional Feature Engineering

With more analysis of statistical data and further research on conventions, we could have added more features to help improve our predictions

Future Applications

Feature Importance

Determine which features have the biggest impact on winning DPOY

Build New Models

Build models to identify the change in a player's defensive ranking when improving the identified features

Leverage Findings

Use our findings to help players improve their defensive game

THANK YOU!