ALL-NBA TEAMS PREDICTION 2022-23

MKT681 Sports Analytics Passion Project

Zayn Sui & Cindy Zhang

Introduction and Background Information

Basketball is one of the most popular sports in the world, with millions of fans tuning in every year to watch the NBA (National Basketball Association) games. One of the most anticipated events of every NBA season is the selection of the All-NBA teams, which recognizes the top players in the league. The All-NBA team is a prestigious honor that reflects a player's hard work, talent, and contribution to their team's success. With only 11 games left for each team in the regular season this year, the discussion about the prediction of All-NBA teams are getting hotter.

The All-NBA team is an annual award given by the NBA to the best players in the league. The team is selected by a panel of sportswriters and broadcasters, who vote for the players based on their regular season performance. The All-NBA team is comprised of three teams, each consisting of five players: two guards, two forwards, and one center.

The selection of the All-NBA team has a significant impact on a player's career. Players who make the team are not only recognized as the best in the league, but they also receive a substantial financial incentive. Given the high stakes of making the All-NBA team, players strive to perform at their best throughout the regular season. Fans and experts also closely follow the All-NBA selection process, as it provides insight into which players are considered the best in the league.

In this project, we will analyze the top players statistics in the league, build models based on historical data, and predict who will make the All-NBA teams for the 2022-2023 season.

Problem Understanding

The primary objective of this project is to predict whether a basketball player will be selected in the All-NBA teams. To determine the three All-NBA teams, the voting results are ranked. The guards and forwards with the most two votes, and the center with the most votes, are selected for the All-NBA first team, and so on. However, the total number of votes varies from year to year. For example, the total number of votes for season 2021-22 is 500, the total number of votes for season 2011-12 is 595. Therefore, the target variable for this project will be the share of votes a player receives. Players who receive a higher share of votes are more likely to be selected for the All-NBA teams.

Data Understanding

Player performance evaluation is a complex process that takes into account various factors. This includes individual performance, which is evaluated based on basic statistics such as points, rebounds, assists, and advanced statistics such as player efficiency rating (PER) and win shares per 48 minutes (WS/48). In addition to individual performance, the player's contribution to the team is also considered, which is reflected by the team's winning percentage.

We used Pandas, Beautiful Soup, and the nba-api libraries to collect the necessary data for this project. The primary source of our data was https://www.basketball-reference.com. To collect the All-NBA teams data, we scraped data from the 1988-89 season, when the number of All-NBA teams increased to three for the first time, up to the last season (2021-22). Additionally, we separately collected NBA player statistics for the current season. However, due to project time constraints, the data for the 2022-23 season is only available up to March 19th, 2023.

There are several criteria for being selected for the All-NBA teams. In our dataset, we typically calculate statistics as an average value. However, if a player does not play enough games or plays only during "garbage" time, their statistics may be similar to those of star players. Therefore, we only consider players who have played at least 60% of the total game time and those who play for at least 30 minutes per game. These are the lowest percentages historically that an All-NBA team player has played.

Modeling

Considering the problem we are solving is a regression problem. We applied 6 popular regression models: Supper Vector Regression, Random Forest Regressor, LightGBM, XGBoost Regressor, AdaBoost Regressor and CatBoost Regressor, to predict the target variable: the share of votes a player will get.

We used the Random Forest Regressor to improve stability and reduce variance by applying the bagging technique. This technique helps to increase the robustness of the model by combining multiple decision trees and their predictions. Boosting, on the other hand, is a technique that involves sequentially adding weak regression models to the ensemble, with each model focusing on the examples that the previous model predicted poorly. This approach can improve the

accuracy and reduce the bias of the model, as the ensemble learns from its mistakes over time. By combining these techniques, we can build a more accurate and robust regression model that is better able to handle complex datasets with high variability.

In this study, the training data consists of 80% of the players who received at least one vote from the 1988-99 season to the 2021-22 season. Similarly, the validation data also consists of 20% of the players who received at least one vote during this time period. The test data, on the other hand, is comprised of players who have played at least 43 games and have an average playtime of 30 minutes per game during the current season.

The evaluation metrics used in this study are the root mean squared error (RMSE) and the coefficient of determination (R^2). RMSE measures the average difference between predicted and actual values and is more sensitive to outliers. On the other hand, the coefficient of determination measures the proportion of the variation in the target variable that is explained by the regression model. It provides a useful measure of how well the model fits the data and can facilitate comparisons between different models. In general, a lower RMSE or a higher R^2 indicates a better-performing model.

Results

After fitting the training data and predicting the validation data for each model, we achieved the summary evaluation.

Model	RMSE	R^2
Random Forest	0.1434	0.7845
SVM	0.1529	0.7552
LightGBM	0.1335	0.8134
XGBoost	0.1391	0.7973
CatBoost	0.147	0.7738
AdaBoost	0.1625	0.7235

The LightGBM model achieved the lowest RMSE and the highest R^2, indicating superior performance compared to the other models. In contrast, the AdaBoost model had the worst performance. We fitted the test data to players who met the minimum number of games played

and minutes per game conditions for the current season, and the resulting prediction results of each model are presented below.

Random Forest					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.7646	Damian Lillard	0.4706	Ja Morant	0.4082
Donovan Mitchell	0.4949	Shai Gilgeous-Alexander	0.4779	Stephen Curry	0.3823
Giannis Antetokounmpo	0.5936	Jimmy Butler	0.4502	Jaylen Brown	0.1971
Jayson Tatum	0.5116	LeBron James	0.2832	Julius Randle	0.1786
Nikola Jokić	0.8122	Joel Embiid	0.8088	Domantas Sabonis	0.2618

SVM					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.6649	Damian Lillard	0.4626	Stephen Curry	0.2953
Shai Gilgeous-Alexander	0.5399	Donovan Mitchell	0.3264	James Harden	0.2856
Giannis Antetokounmpo	0.7155	Jimmy Butler	0.3595	Julius Randle	0.2795
Jayson Tatum	0.5356	LeBron James	0.2843	DeMar DeRozan	0.1799
Joel Embiid	0.9379	Nikola Jokić	0.8473	Anthony Davis	0.3883

LightGBM					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.6745	Shai Gilgeous-Alexander	0.4497	Damian Lillard	0.3712
Donovan Mitchell	0.571	Ja Morant	0.4242	James Harden	0.3531
Giannis Antetokounmpo	0.72	Jimmy Butler	0.5305	DeMar DeRozan	0.1889
Jayson Tatum	0.6379	LeBron James	0.2767	Julius Randle	0.1703
Nikola Jokić	0.8761	Joel Embiid	0.7835	Domantas Sabonis	0.3112

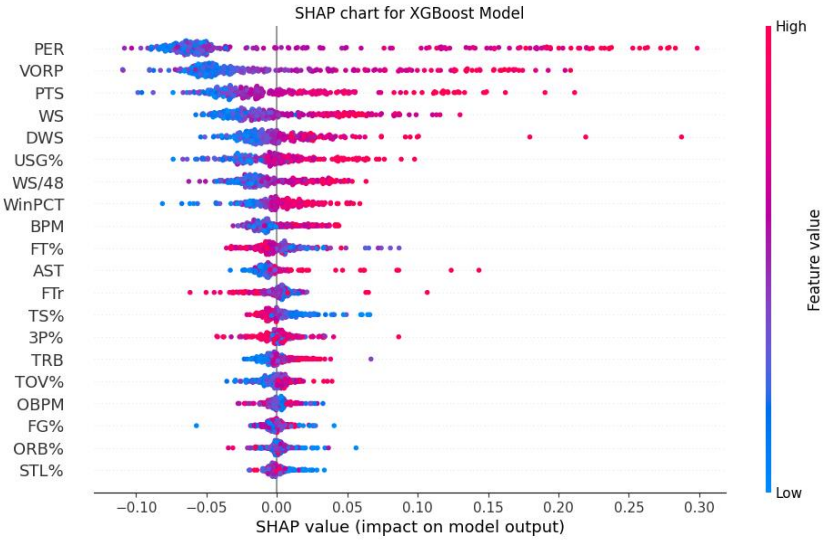
XGBoost					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.8612	Damian Lillard	0.5451	James Harden	0.3211
Donovan Mitchell	0.6548	Stephen Curry	0.4505	Shai Gilgeous-Alexa	0.3196
Giannis Antetokounmpo	0.5291	Jayson Tatum	0.3334	LeBron James	0.3077
Jaylen Brown	0.3448	Jimmy Butler	0.3079	Julius Randle	0.1569
Joel Embiid	0.9159	Nikola Jokić	0.8844	Domantas Sabonis	0.4081

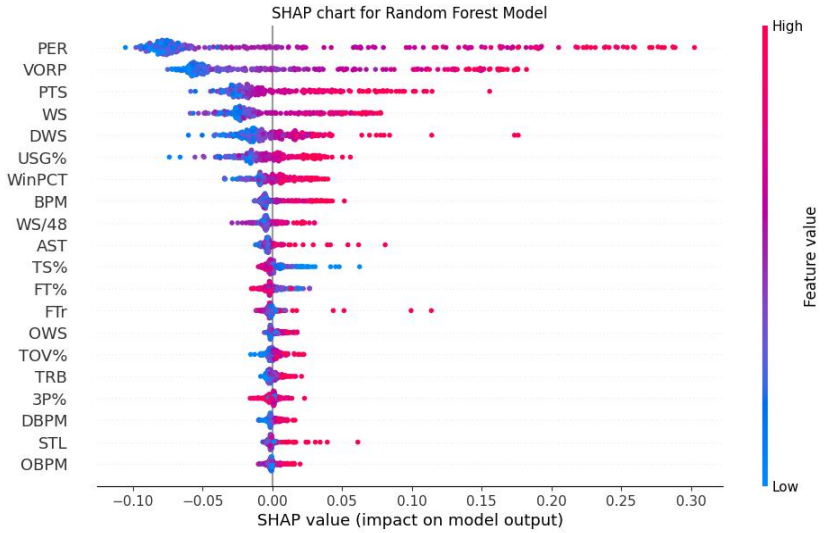
CatBoost					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.7646	Ja Morant	0.4706	Stephen Curry	0.4082
Shai Gilgeous-Alexander	0.4949	Donovan Mitchell	0.4779	Damian Lillard	0.3823
Jayson Tatum	0.5936	Jimmy Butler	0.4502	Julius Randle	0.1971
Giannis Antetokounmpo	0.5116	LeBron James	0.2832	Jaylen Brown	0.1786
Joel Embiid	0.8122	Nikola Jokić	0.8088	Domantas Sabonis	0.2618

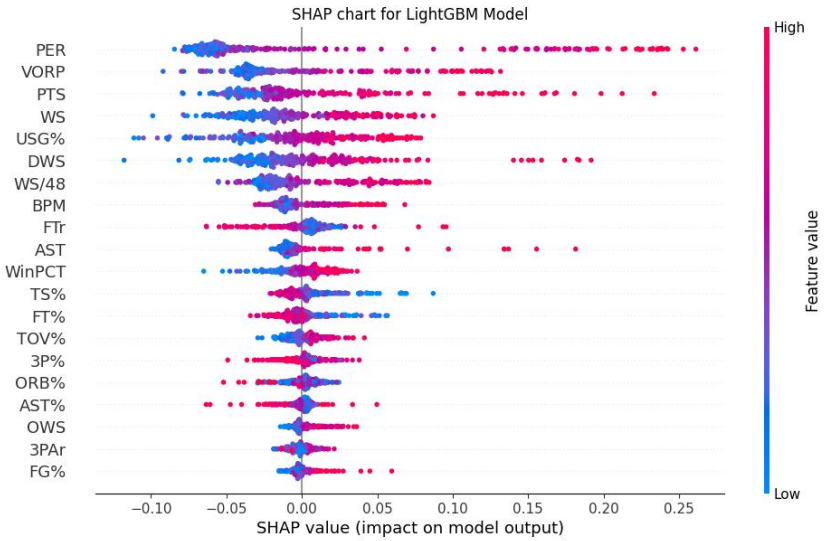
AdaBoost					
All-NBA first team		**All-NBA second team**		**All-NBA third team**	
Luka Dončić	0.6147	Damian Lillard	0.4903	Ja Morant	0.4835
Shai Gilgeous-Alexander	0.5058	Donovan Mitchell	0.4903	Stephen Curry	0.4268
Giannis Antetokounmpo	0.6069	Jimmy Butler	0.4401	Lauri Markkanen	0.2601
Jayson Tatum	0.5484	LeBron James	0.4227	Julius Randle	0.1825
Joel Embiid	0.8062	Nikola Jokić	0.7475	Domantas Sabonis	0.4712

Based on the results, it is evident that Luka Dončić and Giannis Antetokounmpo consistently secure spots in the first team. Meanwhile, Jimmy Butler consistently ranks in the second team, and Julius Randle holds a steady position in the third team. Jayson Tatum makes an appearance in the first team five times, while Joel Embiid appears four times and Nikola Jokić twice. Domantas Sabonis can be found in the third team center position on five occasions. Additionally, either Donovan Mitchell or Shai Gilgeous-Alexander occupies a guard position in the first team. Lastly, LeBron James is typically found in either the second or third team, and Damian Lillard and Stephen Curry can also be found in either the second or third team for most times.

We used the SHapley Additive exPlanations technique to determine the features that were contributing the most to the model's prediction and their impact on the output. The SHAP charts presented below show that the most important features were player efficiency rating (PER), Value over Replacement Player (VORP), points per game (PTS), Win Shares (WS), Defensive Win Shares (DWS), and Usage Percentage (USG%). In summary, these features had the greatest influence on the model's prediction.


SHAP chart for XGBoost Model
PER
VORP
PTS
WS
DWS
USG%
WS/48
WinPCT
BPM
FT%
AST
FTr
TS%
3P%
TRB
TOV%
OBPM
FG%
ORB%
STL%
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
SHAP value (impact on model output)
High
Feature value
Low


SHAP chart for Random Forest Model
PER
VORP
PTS
WS
DWS
USG%
WinPCT
BPM
WS/48
AST
TS%
FT%
FTr
OWS
TOV%
TRB
3P%
DBPM
STL
OBPM
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
SHAP value (impact on model output)
High
Feature value
Low


SHAP chart for LightGBM Model
PER
VORP
PTS
WS
USG%
DWS
WS/48
BPM
FTr
AST
WinPCT
TS%
FT%
TOV%
3P%
ORB%
AST%
OWS
3PAr
FG%
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
SHAP value (impact on model output)
High
Feature value
Low

Conclusion

In conclusion, we used several machine learning models to predict the All-NBA teams for the 2022-23 season. Our models were trained on player statistics from previous seasons and evaluated on test data from the current season. The LightGBM model performed the best, achieving the lowest RMSE of 0.13 and the highest R^2 of 0.81. Using the SHapley Additive exPlanations technique, we identified the key features that influenced the model's prediction.

Based on our modeling result, we predict that the All-NBA first team for the 2022-23 season will include Luka Dončić, Donovan Mitchell, Giannis Antetokounmpo, Jayson Tatum, and Joel Embiid. Second team will be Shai Gilgeous-Alexander, Damian Lillard, Jimmy Butler, LeBron James, and Nikola Jokić. Third team will be Stephen Curry, James Harden, Julius Randle, DeMar DeRozan or Jaylen Brown, and Domantas Sabonis.

While our models have demonstrated good performance, we acknowledge that there is always potential for improvement. For example, if a player makes significant improvements compared to previous seasons, sets historical records, or hits a number of game-winning shots during the season, these factors may not have been fully captured by our current models. Furthermore, when it comes to determining the final All-NBA team selections, a player's on-court performance is not the only consideration. Their off-court conduct may also influence the final votes. Following the horrific event, Ja Morant may face difficulty in being selected for the All-NBA teams this season.