Predicting All-NBA Team Awards With Machine Learning

The Problem

At the end of every NBA (National Basketball Association) season, 3 All-NBA teams, composed of the 15 best players in the league, are selected by voters. This award is one of the highest honors a player can receive in the NBA. As every season nears its end, NBA fans like to speculate and predict which few players will make the distinguished All-NBA teams. Through this project, we intend to use historical data and statistics to find out who really deserves to be All-NBA.

Gathering Data

The first step in this process was to decide what data our machine learning model will learn from and where to get it. We decided to use the public nba_api library to gather our data as it allowed us to quickly and effortlessly retrieve large amounts of relevant data.

The data we decided to gather included all basic counting stats (points, rebounds, assists, etc.) and common NBA advanced stats (offensive rating, true shooting percentage, player impact estimate, etc.) for every player in the league. Furthermore, we needed to gather stats for every team in the league, as some of the additional advanced stats we intended to calculate later on would require data related to a given player's team.

We then had to get this data for every season in a given interval. The nba_api library would only let us retrieve data as far back as the 1996-97 season so our chosen interval was the 1996-97 season to the current 2023-24 season. This limitation to the 1996-97 is not a hindrance to our dataset as there were many in-game rule changes made and changes to how All-NBA teams were selected prior to this season.

Finally, we needed a list of every All-NBA player throughout the NBA's history, which we acquired quite easily thanks to basketball-reference.com.

Cleaning The Data

After we gathered the data we wanted, the first step in cleaning it was to remove irrelevant columns from our dataset. The nba_api library gave us a total of 123 columns of data, most of which contained stats that were useless for this project. We ended up narrowing down our datasets to just 50 columns of stats that are or may be important later on.

We also had to clean the All-NBA teams data retrieved from basketball-reference.com as it was in a different format compared to the rest of our data. This process involved normalizing player names that included accented letters and removing position suffixes that were included at the end of every player’s name.

Additional Advanced Stats

While we already had access to plenty of advanced NBA statistics through the nba_api library, we wanted to calculate two additional advanced stats that are very popular for comparing the best players in the league. The first stat we calculated was win shares (WS), which is described as an estimate of the number of wins contributed by a player according to basketball-reference.com. The second stats we calculated was player efficiency rating (PER), which is described as a per-minute rating of a player's performance.

Both of these advanced statistics are not directly available through the nba_api library, but they can be calculated with the data that was provided. All equations used to calculate WS and PER were provided by basketball-reference.com.

Finalizing The Data

Our final step before making predictions was to finalize and set up all of our data for the machine learning model. We did this by combining our clean datasets (player stats, additional advanced stats, and All-NBA teams) into one final player stats dataset. Then, we reduced the number of columns in our data even further to only include stats necessary for our model and for us to understand any results.

Finally, we separated the data by seasons to create three files, one containing the seasons from 1996-97 to 2021-22 which would be our training data for the machine learning model. The next file contains just the 2022-23 season, which we will make predictions on with our model to test its accuracy on a single season. The third and final file contains the data for 2023-24 season, which is the currently running NBA season and does not contain All-NBA team information yet as the awards haven’t been announced at the time of this report. We will still make predictions on this season just for fun!

Machine Learning

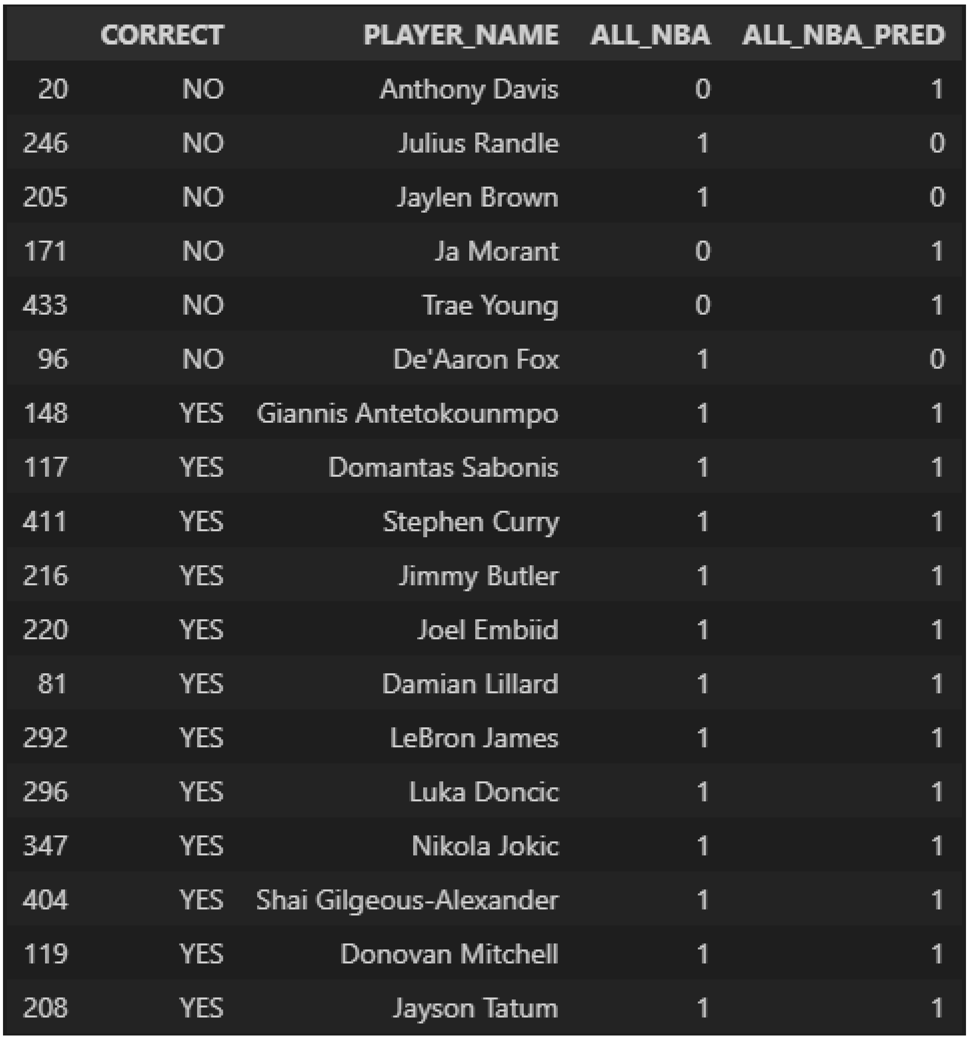
The machine learning model that we decided to use to predict All-NBA players was a Random Forest Classifier. One of the reasons we chose this model is because it handles imbalanced classes quite well. This is important as there are only 15 All-NBA players in a given season out of upwards of 500 NBA players. Additionally, the

Random Forest model is able to handle the non-linear relationships between NBA player statistics without the need of a standardized scaler. This model also scales well with the amount and complexity of data that we have.

We trained the model with a 75/25 training/test split of the training dataset created in the finalizing data step. We then fit the Random Forest Classifier to the training data. Once the model was trained on over 25 years of NBA data, we were able to use it to predict All-NBA players for the 2022-23 season to measure its accuracy when used on a single-season dataset.

Findings and Analysis

Below are the results of our models predictions on the 2022-23 season compared with the actual All-NBA team players.



	CORRECT	PLAYER_NAME	ALL_NBA	ALL_NBA_PRED
20	NO	Anthony Davis	0	1
246	NO	Julius Randle	1	0
205	NO	Jaylen Brown	1	0
171	NO	Ja Morant	0	1
433	NO	Trae Young	0	1
96	NO	De'Aaron Fox	1	0
148	YES	Giannis Antetokounmpo	1	1
117	YES	Domantas Sabonis	1	1
411	YES	Stephen Curry	1	1
216	YES	Jimmy Butler	1	1
220	YES	Joel Embiid	1	1
81	YES	Damian Lillard	1	1
292	YES	LeBron James	1	1
296	YES	Luka Doncic	1	1
347	YES	Nikola Jokic	1	1
404	YES	Shai Gilgeous-Alexander	1	1
119	YES	Donovan Mitchell	1	1
208	YES	Jayson Tatum	1	1

As seen in the table, our model correctly predicted 12 out of 15 All-NBA players resulting in an accuracy score of 80%. At first this may not seem very impressive considering how few players make up the All-NBA teams. But with further analysis, our model may be more accurate at selecting the 15 best players in the league (at least statistically) than it seems.

Principal Component Analysis

To gain more insight on our models predictions, we performed a principal component analysis (PCA). The resulting variance ratio showed that the first principal component had much higher variance than the second principal component, making it more impactful to our models predictions. We also learned what the most important features of each principal component was.

First principal component

The top 5 features in order of decreasing variance were: Points, Player Impact Estimate (PIE), Player Efficiency Rating (PER), Rebounds, Assists.
All of the features seem to be fairly weighted with Points having the highest variance. The only exception for the first component was Defensive Rating which had a much lower variance compared to the others.

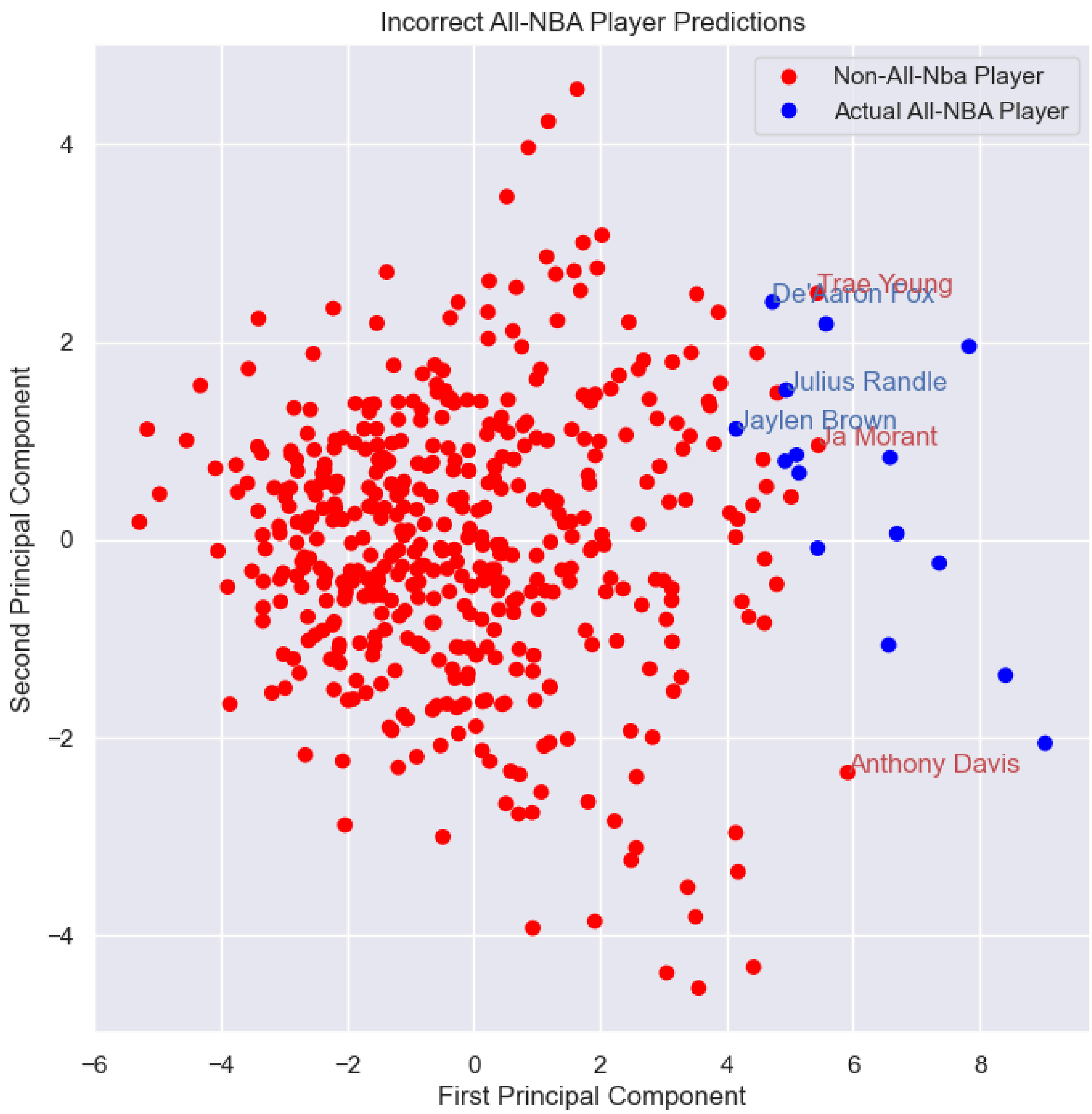
Second principal component

The top 5 features in order of decreasing variance were: Defensive Rating, Usage Percentage, Assists, Points, PER.
The second principal component valued Defensive Rating the most, which could explain the exception in the first component. Usage Percentage also had a much higher variance than the rest of the features.

Both principal components had higher variances for Points, Assists, and PER which indicates that these are some of the most important stats for selecting All-NBA players.

Visualization and Further Analysis

The graph below shows the incorrect predictions that our model made. The player names in red are players that the model believed to be All-NBA level, but were not selected to an All-NBA team. The player names in blue are players that the model believed were not All-NBA level, when they actually were.



Our model incorrectly predicted that Trae Young, Ja Morant, and Anthony Davis were All-NBA level players. This is quite interesting as these three players were widely considered as some of the biggest All-NBA team snubs of the 2022-23 season, indicating that our model may be more accurate than we originally thought.

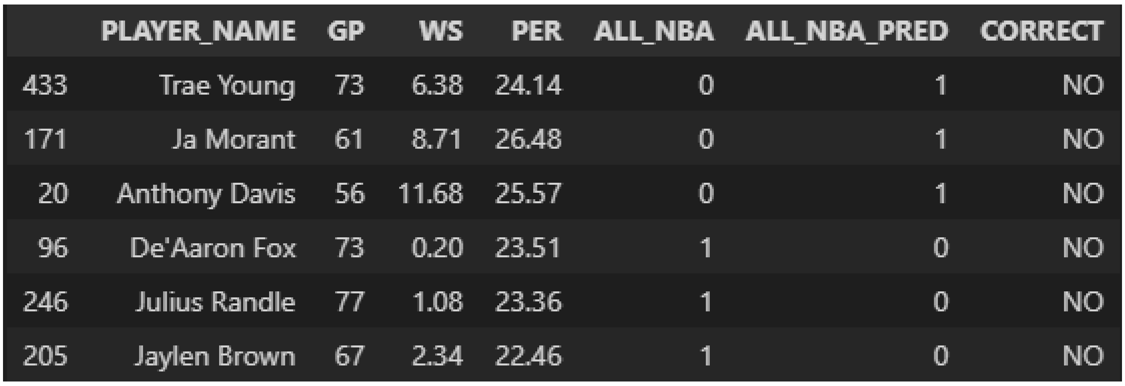
From this graph we can see that these players were clearly weighted more positively on the x-axis compared to the other 3 actual All-NBA players (De'Aaron Fox, Julius Randle, and Jaylen Brown). The reason that these players weren't actually selected as All-NBA players is likely due to various influences that our model didn't account for.

One of these being that, in reality, players aren't selected by an algorithm or purely based off of statistics. Instead, a group of 100 sportswriters and broadcasters get to vote for players they personally believe deserve the award. These voters may use information that our model isn't aware of to inform their decisions.

For example, our model doesn't particularly value the performance of a player's team. For example, Jaylen Brown's team, the Boston Celtics, had 16 more wins than Trae Young's team, the Atlanta Hawks. This may have been a key factor that led to Jaylen Brown being an All-NBA player over someone like Trae Young.

In addition, our model doesn't know about games played. Games played is largely considered the reason why Anthony Davis was snubbed as he only played in 56 games compared to actual All-NBA forward Julius Randle's 77 games played (who our model did not believe to be All-NBA level). Ja Morant also played less games than the others.

Another notable difference between these six players is that the three players that our model selected all had higher Player Efficiency Rating and Win Shares than the 3 actual All-NBA players. This makes sense as our principal component analysis revealed that Player Efficiency Rating was particularly important to our model.



	PLAYER_NAME	GP	WS	PER	ALL_NBA	ALL_NBA_PRED	CORRECT
433	Trae Young	73	6.38	24.14	0	1	NO
171	Ja Morant	61	8.71	26.48	0	1	NO
20	Anthony Davis	56	11.68	25.57	0	1	NO
96	De'Aaron Fox	73	0.20	23.51	1	0	NO
246	Julius Randle	77	1.08	23.36	1	0	NO
205	Jaylen Brown	67	2.34	22.46	1	0	NO

Overall, our model's mistakes seem to be somewhat reasonable and in the eyes of many NBA fans, our model may even be more accurate than the accuracy score of 80% would suggest.

Limitations

One of the main limitations with our machine learning model was that we couldn't force it to always select exactly 15 All-NBA level players. While testing and modifying our model, we would often get results that predicted more or less than 15 players. Fortunately, the random seed we ended up on happened to predict exactly 15 players.

Something we wanted to add but were unable to in a reasonable amount of time was the positional limit rule. Up until the current NBA season, All-NBA teams were always composed of one center, two forwards, and two guards. Unfortunately it is far too difficult to determine any given player's position accurately, especially in the modern NBA, as they can play multiple different positions throughout a season. This rule may have caused mistakes in our models predictions as it would often select more guards than forwards when in reality it should be selecting an equal amount of both. Interestingly, this rule has been removed in the current NBA season, now players with the most votes are selected to All-NBA teams, regardless of position.

This NBA season also introduced complicated minimum games played requirements for various NBA awards, including All-NBA teams. We could have applied this rule to our predictions, but all the historical data that our model was trained on does not abide by this new rule, so we decided against it.

Project Experience Summary

Mikael Kaas - 301457764

- Gathered, cleaned, and prepared large amounts of data for a machine learning model
- Generated predictions for All-NBA team award winners for the 2022-23 and 2023-24 NBA seasons using a Random Forest Classifier machine learning model
- Performed Principal Component Analysis on the model and visualized the results to make further analysis on the relationship between NBA player statistics and All-NBA team selections