Projecting Future All-NBA Probabilities

Joe Siwinski
October 1, 2019

Introduction:

My objective is to build a model to project the probability of a current NBA player making an All-NBA team at any point in their career.

To solve this problem, I split the project into different stages.

Collect Data

- The first stage is to collect the appropriate data for All-NBA probability creations. This will be a combination of both traditional box score statistics and advanced statistics such as VORP (Value Over Replacement Player).

Boosting Model for All-NBA Probs.

- The second stage is to actually train a classification model using past NBA seasons data to project probabilities of All-NBA selections. As I will mention later, my model of choice to accomplish this is a binary logistic boosting model using xGboost in R.

Projecting Careers with Historical Similarity Scores

- The third stage is to come up with forecasted statistics of each player in the NBA. I will use the Euclidean Distance based Similarity score formula to approximate player statistics based on past similar players in the NBA.

Applying Boosting Model to Projected Careers

- The fourth stage is to apply the All-NBA probability model made in Stage #2 to the forecasted statistics in Stage #3. This will accomplish the goal of mapping out future probabilities of making an All-NBA team for each player in the NBA.

Visualization of Findings

- The last stage is focused on presentation of my findings. I will utilize Shiny in R to create an interactive line graph that displays each player's projected probabilities by future season. This will give the user a way to visualize the results of my project and compare players future outlooks to one another.

For the entire project I will be utilizing R for every stage. My code can be seen in the attached R scripts.

Stage 1 – Collecting Data:

In my past knowledge and experience with watching the NBA, I believe that the all-NBA teams votes are based on a mixture of individual box score statistics, individual advanced statistics, as well as the strength of the team the player is on (ex: Win-Loss record). Unfortunately, because I do not know what team each player will be on in later years of their career, I will only focus on the individual player statistics (traditional and advanced) portion of voting.

The statistics that I used to project probability of making an All-NBA team are as follows. I scraped these statistics utilizing the "ballr" package in R.

Table 1: Statistics used for Boosting Classification Model

Variable	Description
Position	Position of Player (Guard, Forward, Center)
PPG	Points per Game
APG	Assists per Game
RPG	Rebounds per Game
OWS	Offensive Win Share
DWS	Defensive Win Share
VORP	Value over Replacement Player
USG %	Usage Rate of Player

Stage 2 – Boosting Model for All-NBA Probabilities:

My modeling technique of choice is a boosting model utilizing xGboost in R. This will be a binary logistic model that predicts the probability of a player either making an all-NBA team or not (1 = make All-NBA team, 0 = not on All-NBA team). I will use k-fold cross validation and use a grid search to find optimal hyperparameters.

In addition, I limited the data used in the modeling process to only years after the 1988-1989 season. I made this decision because in this season, there were 3 All-NBA teams as compared to only 2 All-NBA teams in years before. Another change I made to the data was to limit the players included in the modeling process to players who played over a half season (41 games) to restrict players who might have high averages but only from a small sample size of games.

Further, I scaled each player's statistics by each season. I did this to include relative overall league changes in statistics year to year. Let's say one year all players averaged generally less points per game than another season of players. (Recent league years have higher scores due to increase of pace of play). This scaling process helped improve model accuracy by taking these league wide changes into account.

I included three interactions in my model that improved model performance. These were interactions between DWS and Centers, RPG and Centers, as well as APG and Guards.

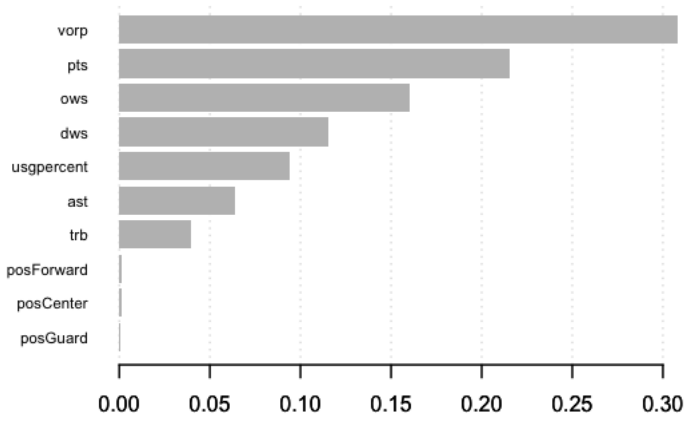
The results from the grid search after feature engineering were the following optimal hyper parameters:

Model 1: All-NBA Probability Optimal Boosting Parameters

Test Log Loss	Train Log Loss	Sub Sample Rate	Col Sample Rate	Depth	Eta	Current Min Child
0.05	0.023	0.8	0.6	4	0.10	1

Feature importance of the boosting model can be seen in the graph below:

Figure 1: Feature Importance in Boosting Model



After training a boosting model with the above tuned hyperparameters, I was ready to apply it to future player seasons to obtain All-NBA probabilities.

Stage 3 – Projecting Careers with Historical Similarity Scores:

In order to apply my All-NBA boosting model to future seasons, I needed to forecast how each player would perform in future years. I would need a value for each of the statistics listed in Table 1 to generate these All-NBA probabilities.

To describe the process of career trajectory building that I used, I will run through the example with Donovan Mitchell. After calculating Euclidean distances between players in my database (Based on individual stats in boosting model as well as others such as 3 point attempt percentage and true shooting percentage), I filtered the top 50 most similar players that were the same age as Mitchell (22). With these players filtered, I then selected every season that these similar players

had after their 22 year old season. I would use the future season trends of the similar players as a way to forecast how Mitchell's statistics in his subsequent seasons after his 22 year old season. So for instance, the most similar player to 22 year old version Donovan Mitchell is Ben Gordon from Gordon's 22 year old season. So, the change in Gordon's PPG from his 22 year old season to his 23 year old season will contribute to Mitchell's forecasted PPG and other statistics in his next year in addition to the 49 other players filtered. This process would continue for all stats and for all seasons played by Mitchell.

I decided to use a weighted average based on similarity scores calculated from the following distance based similarity formula.[1]

$$\frac{1}{1 + d(p_1, p_2)}$$

I used weighted averages to give more of an impact from the most similar players in comparison to the less similar players in the selected 50. So in this case, Ben Gordon would have the highest contribution to Mitchell's forecasts while similar players like Lou Williams (40th most similar player) would have a smaller contribution.

Deciding Retirement:

While retirement is something that is hard to project, I made a rule to force a player to retire when less than or equal to 10% of the top-50 similar players were still playing. In Mitchell's case, 7 out of the 50 similar players were still playing at age 37. At 38, only 5 of the players were still playing thus making Mitchell's 37 year old season his last.

Exceptions:

Two exceptions that I made were with Kristaps Porzingis and Jakarr Sampson. I used their 2017-18 statistics and forecasted from this season as Porzingis did not play in the 2019 season and Sampson is not a 20 PPG player (only 4 games played in 2019).

Process Illustration:

To simplify this process further, I will illustrate how my code worked and how each subsequent future season stats would be calculated for each player. I will stick with the Donovan Mitchell example to display this. The process is drawn out on the next page.

[1] d(p1, p2) = Euclidean Distance between player of interest (Mitchell) and another player

DONOVAN MITCHELL CAREER FORECAST

Actual 2018-19 Stats							
Age	**PPG**	**APG**	**RPG**	**OWS**	**DWS**	**VORP**	**USG**
22	23.8	4.2	4.1	1.3	3.7	1.7	31.6%

Weighted Avg. Changes in Statistics

0.4	0.3	0.3	1.0	0.4	0.6	-0.6%

Forecasted Stats							
Age	**PPG**	**APG**	**RPG**	**OWS**	**DWS**	**VORP**	**USG**
23	24.2	4.5	4.4	2.3	4.1	2.3	31.0%

Process continues for each subsequent season until Mitchell Retires (10% Rule)

Stage 4 – Applying Boosting Model to Projected Careers:

With forecasted statistics for each player, I now was able to apply my boosting classification model to come up with probabilities of each player making an All-NBA team in future seasons. I have included selected players forecasted All-NBA selection probabilities for their remaining years in the NBA.

Player	Career	'20	'21	'22	'23	'24	'25	'26	'27	'28	'29	'30	'31	'32	'33	'34	35	Projected
Doncic	0	74%	93%	95%	95%	96%	95%	84%	92%	89%	79%	64%	65%	12%	4%	2%		10.4
Curry	6	87%	50%	13%	6%	2%	0%	0%	0%	0%								7.6
Towns	1	87%	89%	79%	87%	64%	64%	62%	28%	4%	2%	1%	0%	0%	0%	0%	0%	6.7
Irving	2	66%	46%	33%	6%	4%	2%	1%	1%	0%	0%	0%	0%					3.6

The likelihood of each player getting the most All-NBA selections from here on out can be seen below. To come up with the probabilities, I assumed a binomial distribution with the above probabilities. I utilized R to come up with probabilities for every combination of scenarios and combined these into the following table.

Player	Prob. Of Most
Doncic	99.2215700%
Towns	0.1451214%
Irving	0.0000006%
Curry	0.0000001%
Tie	0.6333116%
Total	100.00000%

The results of my project show just how special Luka Doncic is as a player. It appears that Doncic will be a top NBA player for many years to come.

Stage 5 – Visualization of Findings:

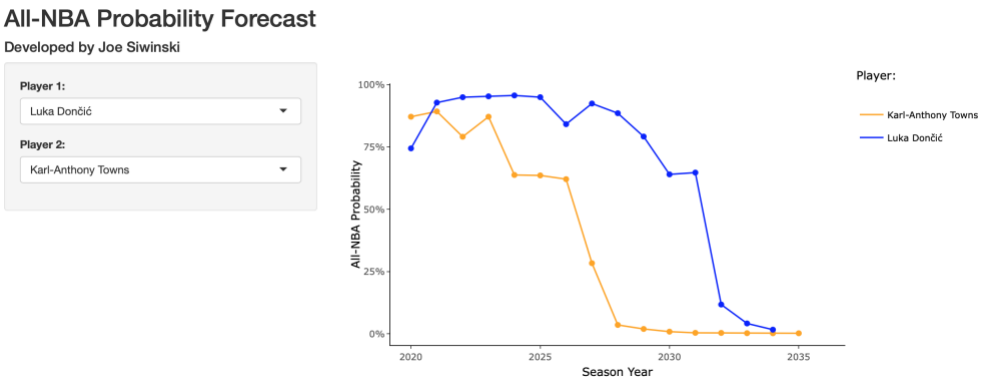
For my visualization, I have developed an interactive and dynamic app using Shiny, ggplot2, and plotly in R. The app allows the user to input any current player into the select box and see that player's future All-NBA probabilities by season. The app also allows the user to compare multiple players chances at All-NBA selections by year.

I really wanted to allow the user to see a player's future All NBA probabilities instantly upon inputting the desired player's name. Additionally, I wanted to allow the user to compare players to gain a better insight on a certain player's future outlook as compared to another player of interest.

I have included a screenshot displaying two selected players and have also provided the link to the full application on the following page:

Link to see full app:

https://joesiwinski.shinyapps.io/AllNBAProjections/



Future Improvements:

In the future, I can make a few improvements to better my projections. These can be seen below:

1) Rookie projections
    a. I can add in a rookie projection model based on a player's college stats to include these projections. These projections could give teams a better idea of what prospects will do well in the NBA.

2) Similarity improvements
    a. In addition to the variables that I used to compare players, I can add in shot location frequency stats to further improve player projections.

3) Injury Simulations
    a. To improve the projections further, I can add an injury feature to the projections that also includes the durability of each player. This would effect older players as well as players with injury history (ex: DeMarcus Cousins).