Constructing My All-Time NBA Lineup Through K-Means Clustering, Part 2

By Nikhil Sharma

Nell Redmond, The Associated Press

If you haven't read my first article on the topic, check it out here!

A few months ago, I wrote an article where I used *k*-means clustering to help choose five NBA players from 1980 to the present to constitute an all-time starting lineup.

Mostly, the feedback was great, and people were able to understand the process I went through to choose my starting 5. However, I also got some negative feedback, and would like to address some of that here. While I received the typical "lol Draymond" comments, I also got some constructive criticism on my methodology that certainly calls for response.

I got a few comments on how my lineup does not factor in era strength. While it is true that the strength of competition vastly varies throughout different eras, I would like to think that by only using All-Stars from the 1979-80 era to the present, I eliminate a good deal of the confoundment across eras. Though some mid-to-lower tier players might have had more success if they played in a different decade (e.g., Jahlil Okafor), I am of the opinion that elite players are elite, regardless of the era they play in. To say that my lineup of Magic Johnson, Michael Jordan, Larry Bird, Draymond Green and Shaquille O'Neal would get "rekt" because era strength was not considered is, frankly, a bit unreasonable. I understand that a lot of those older players played in a time where the league was just not as talented, fast, etc., but saying players like Magic and Jordan would not be as good had they played today completely ignores the superior skill, talent and work ethic these players possessed that made them the legends that they are.

Similarly, some negatively reacted to my use of advanced stats, since "it's easy to play to maximize an analytic after it is created", and thus invalidating the use of advanced analytics on players from different periods of the game. However, if we look at how some of these advanced statistics are calculated, it is clear that players and data from all eras are carefully considered when regressing various statistics to produce the target advanced statistic.

Still, let's look at some of the top 10 players in my dataset for some popular advanced statistics.

Win Shares:

Michael Jordan	1988
Michael Jordan	1996
LeBron James	2009
Michael Jordan	1991
David Robinson	1994
Michael Jordan	1989
LeBron James	2013
Kevin Durant	2014
Michael Jordan	1990
Kevin Durant	2013

Value Over Replacement Player:

Russell Westbrook	2017
Michael Jordan	1989
Michael Jordan	1988
LeBron James	2009
LeBron James	2010
David Robinson	1994
LeBron James	2008
Michael Jordan	1990
Chris Paul	2009
Kevin Garnett	2004

Offensive Box Plus-Minus:

Stephen Curry	2016
Russell Westbrook	2017
Stephen Curry	2018
Michael Jordan	1988
Michael Jordan	1989
Tracy McGrady	2003
LeBron James	2010
Michael Jordan	1990
James Harden	2018
Stephen Curry	2015

Defensive Box Plus-Minus:

Ben Wallace	2003
Ben Wallace	2004
Ben Wallace	2006
David Robinson	1992
Mark Eaton	1989
Hakeem Olajuwon	1990
Joakim Noah	2014
Ben Wallace	2005
Hakeem Olajuwon	1993
Andre Drummond	2018

While this is not the most precise way to show this, we can see that for many of the popular advanced metrics, there is a decent amount of variation in players from different time periods when focusing on the cream of the crop.

In a broad sense, I feel that the disagreements here are a matter of individual perspective on whether it is fair to compare statistics over time despite disparities in league strength and playing style. While I generally agree that many of these statistics confound due to time, I think that the structure of my project mitigates that enough to still be considered valid.

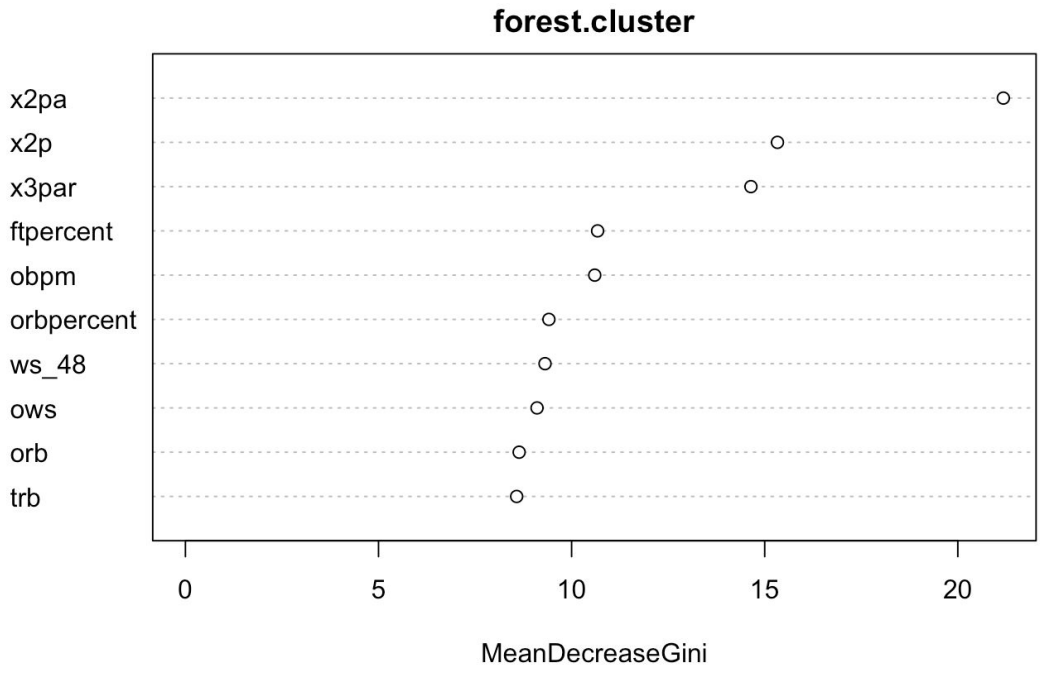
I'd like to switch gears to the suggestions I got about finding more statistical ways to choose players from the clusters. I will admit that I could have at least considered some more math-y methods, even if I still ended up choosing players I saw fit for an all-time lineup.

In my original article, I did list players that were closest to the centroid in each cluster; we can consider players from this group in a possible lineup.

I also used a random forest method to tease out important features that distinguished players in each cluster from players in the entire dataset. For reference:

> To further investigate, I carried out a method recommended by this article to extract feature importance in clusters. First, I constructed a random forest model that predicted the probability of every player belonging to the 7th cluster. Then, using a built-in function in the randomForest library in R, I saw which variables were most important in construction of the model. This showed me that in determining whether a player was in the 7th cluster or not, the two most important variables were 2 Pointers Attempted Per Game and 3-Point Attempt Rate (percentage of field goal attempts that were 3-pointers).

Here is what the plot that shows variable importance in the randomForest library looks like:



To approach lineup construction in a more scientific method, I decided to perform this random forest process on each cluster, then arrange the players in each cluster in descending order of their most important feature, and finally, pick the player with the highest value.

Through the closest-to-centroid and random forest methods, these are the players we end up with:

Cluster	Closest-to-Centroid	Random Forest
1	Eddie Johnson, 1981	Derrick Rose, 2012
2	Charles Barkley, 1995	Hakeem Olajuwon, 1990
3	Marques Johnson, 1986	Moses Malone, 1988
4	Chris Mullin, 1990	Steph Curry, 2016
5	Jim Paxson, 1983	George Gervin, 1980
6	Chris Bosh, 2014	Shawn Marion, 2007
7	Peja Stojakovic, 2002	Tim Hardaway, 1991

To be blunt, these groups of players do not stack up to the starting five that I assembled. However, lineups created from both collectives would be incredibly fun.

From the Closest-to-Centroid group, I would probably pick Marques Johnson, Chris Mullin, Peja Stojakovic, Charles Barkley and Chris Bosh. This would be an electrifying lineup; we would have the first ever point-forward (and fellow Bruin) in Johnson running the show, two elite sharpshooters on the wing in Mullin and Stojakovic, one of the greatest and most uniquely talented players of all time in Barkley at the 4 and an underrated, elite defensive big who could stretch the floor and occasionally make a nifty move in the post in late-career (*sad react*) Bosh. This team would also have a distinctive defensive identity, as the shortest player here is Barkley at 6'6". Unfortunately, Mullin and Stojakovic would almost certainly have trouble defending on the wing; despite their size and length, neither were known as the most lockdown guys in their days.

From the Random Forest group, I would pick out Steph Curry, Derrick Rose, George Gervin, Shawn Marion and Hakeem Olajuwon. On the offensive end, this lineup would be so ridiculously entertaining to watch. At the guard slots, we would have the greatest shooter to ever grace basketball in Curry and one of the quickest, most explosive athletes to play in the NBA in Rose (ignoring injuries *sad react*). The wings would consist of the smoothest scorer out there in The Iceman and the amazingly athletic Marion, who would compete with Curry on this team for the prettiest jump-shot. And finally, lurking in the paint, we have the crown jewel of this team: Hakeem Olajuwon. As arguably the greatest center in NBA history, Hakeem would make defenders look silly with his silky post-moves on a nightly basis. While this team would be absolutely unstoppable on the offensive end, they would certainly struggle on defense. Despite having the all-time leader in blocks manning the paint (Olajuwon) and a marvelous, versatile defender available to switch on whomever (Marion), this lineup has a pretty subpar defender on

the wing in Gervin. And as for the guards, we might as well put turnstiles on the perimeter if our options out there are Curry and Rose.

Ultimately, mostly for defensive reasons, I would still choose my lineup of Magic Johnson, Michael Jordan, Larry Bird, Draymond Green and Shaquille O'Neal over these more scientifically chosen lineups. Regardless, these lineups were important exercises to consider, and definitely added a bit more of a statistical element to my decision, as they gave me perspective on two numerically defined alternatives to my initial choosing.

This is likely the end of my work on this project, unless I come back a few months later with a random revelation, or decide to respond to a few more Reddit comments. Thank you for indulging me in this long, arduous, basketball-y journey.

But again, *who is scoring on my team?*

If you want to see my code, you can find it here.