ML for MVPs:

Using Machine Learning to Predict the NBA’s Most Valuable Player

Fisayo Omilana

June 5, 2016

This year in the NBA, one of the biggest questions has been the race for MVP and who deserves the award. Multiple players this year have been putting up historic numbers and having extraordinary seasons which has created a rift amongst NBA fans as to who is the clear-cut Most Valuable Player. To name a few in spotlight: Russell Westbrook, James Harden, LeBron James, and Kawhi Leonard. By taking data from previous seasons and flagging the MVP winners of each past season, I hope to answer this question.

I used data from ESPN.com and BasketballReference.com for this entire project. The first database from ESPN I used contains yearly traditional box score player statistics. The second database contains more advanced metrics. The database from Basketball Reference has annual results of the NBA Awards Voting. Specifically, the three links for my sources are:

* <http://www.espn.com/nba/statistics/player/_/stat/scoring-per-game/sort/avgPoints/seasontype/2>
* <http://insider.espn.com/nba/hollinger/statistics>
* <http://www.basketball-reference.com/awards/awards_2003.html>

To extract data from ESPN I utilized a Google Chrome-compatible applet called Scraper. By highlighting the data and right clicking, I scraped a page of data at a time and exported it to Google Sheets. The export followed a loose organizational layout but it was not perfect. Consequently, lots of clean-up needed to be done. Repeat labels had to removed/renamed, ambiguous unique identifiers needed to be redefined, empty cells needed to be filled in with default values, multiple tables needed to be merged into a single all-encompassing spreadsheet, and the MVP Results needed to be manually inputted. For the first couple of seasons, I used the =VLOOKUP() function to merge sheets but then I found more useful add-ons within Google Sheets to speed up the process. In my initial training and testing, I only measured accuracy but soon found out that the number of possibilities for true negatives within my data was confounding the robustness of my model. To counteract this issue and account for positive predictions, I began to use other evaluation metrics in my models (specifically, precision and recall). Using supervised learning to train and test several classifiers on the numerous statistical categories, I found that Random Forests with an input of 10 trees yielded the best overall performance. In one instance, it resulted with a 99% training accuracy, a 92% training precision, a 100% training recall, an 89% testing accuracy, a 60% testing precision, and an 81% testing recall. Because of this performance, I used Random Forests (10 trees) to predict this year’s MVP Race.

Since the results from each run of the model varies randomly, I came up with a sort of tier system to determine a generalized result. The system works by ranking based on how many times a player is picked in the top finishers of the MVP Race for ten sample runs. To break ties between players who were ranked the same number of times, I compared the average rank each achieved for all sample runs. Using this tier system for ten total runs of the Random Forests model, my results were as follows:

LeBron James, 1st place with an average rank of 2.4 for 10 sample runs

Russell Westbrook, 2nd place with an average rank of 3.1 for 10 sample runs

James Harden, 3rd place with an average rank of 2.1 for 7 sample runs

Kawhi Leonard, 4th place with an average rank of 1.3 for 6 sample runs

Anthony Davis, 5th place with an average rank of 2.7 for 4 sample runs

Given my own NBA fandom, the model’s results were reasonable with a slight surprise because the model predicted LeBron James as the 2016-17 Most Valuable Player. This prediction was particularly surprising since he was recently snubbed from the final three candidates for the award. However, without this knowledge, the model is still making a logical prediction given that James is probably one the greatest basketball players to the ever play the game, has been averaging career-high numbers in multiple statistical categories such as rebounding and assists, and was, statistically speaking, the most efficient 20+ point scorer in the league this year. Besides LeBron James, the rest of my results made a lot of sense and might very well end up being the actual result of the MVP Race this reason.

After evaluating my model’s predictions, I became interested in knowing which features had the most impact. By finding these features, I hypothesized that I could essentially determine which specific statistical categories were most important to consider when distinguishing an MVP. This basically involved computing the total difference in my model’s performance with and without each feature. After these calculations, I listed each feature with its respective net effect on the model. Once I found these most impactful statistical categories, I decided to evaluate their trends over time. By breaking up all my data into five-year intervals and taking averages of MVP candidates from each mini-era in the significantly telling categories—Usage Rate; Points Per Game; Player Efficiency Rating; Total Assists; Minutes Per Game; Triple Doubles; & Double Doubles—I created Tableau visualizations of the progression/regression of standards for NBA MVP’s.

full link: <https://public.tableau.com/profile/fisayo.omilana#!/vizhome/NBAMVPStandardsOvertheYears/NBAMVPStandardsOvertheYears>

Over time, most of the numbers were increasing in many of the categories. The data shows that the caliber of play/output among the NBA’s best has grown higher and higher (USG, PPG, PER, AST, & TRIDBL) in the past fifteen years. This idea makes sense especially in today’s fast paced, high scoring game in comparison to the slower, more methodical low-post scoring style of the past. In addition, it is interesting to see how the average number of triple doubles has skyrocketed in the past five years. This is undoubtedly due to players who have recently become infamous for the impressive stat line such as: Russell Westbrook, LeBron James, and James Harden. Conversely, it is also notable that the number of double doubles has decreased. The 2002-2007 season has a considerably larger average number of double doubles among MVPs than the other two mini-eras. Looking at the labeled data for this mini-era, many of the candidates are frontcourt players who regularly tallied double digit scoring and 10+ boards. A big. At the time, these MVP-caliber big men were the focal point of their team’s offense. I inferred that a likely reason these numbers are dropping is because, as mentioned before, the entire league is straying away from the old-fashioned game of dominating the paint. I believe this decline of double-doubles among MVPs is basically representative of the gradual extinction of token post-up NBA centers and power forwards. These once previously invaluable assets are now being replaced by athletic, high-scoring guards as hot commodities of the NBA. Lastly, another declining trend is Minutes Per Game. This is extremely pertinent to today’s game regarding the issue of resting players. Throughout the league, more and more teams are being careful with their star player’s minute totals. Assuming the likely situation that a MVP caliber player’s team is doing well, successful teams tend to rest their best player at certain points in the season because reducing workloads can keep a player fresh, minimize the chances of an injury, and even extend his career in the long run. Also, successful teams are not worried about underperforming while their best player is resting because a possible loss will likely not drastically affect their playoff position. Therefore, for many winning franchises, the long-term benefit of resting a star player greatly outweighs the low-risk outcome of losing one regular season game. As the data shows, this was not the case in the past. It is possible that franchises either did not consider extended minutes as an adverse factor and/or the competitiveness of the league was much more intense where MVP players were needed on the floor more often than not.

One attribute that I feel would have also been a good indicator, but was unable to cleanly account for in the data was a MVP candidate’s team’s win percentage. All in all, it is hard to say who is the *Most* Most Valuable Player of all the award winners since the game of basketball is always changing and each player’s value can very well be time-dependent. Nevertheless, this project truly gave me a glimpse of how difficult it can be to unanimously stand out as the league’s best of the best.

Description of Files Used

2016-17blank.csv -- This file has player statistics but no final results on the MVP Race

2016-17labeled.csv -- Files with the “labeled” tag contain corresponding player names

mvp-trnblank.csv -- This file is a consolidation of data for seasons 2002-2013 for training models

mvp-trnlabeled.csv -- (see above)

mvp-testblank.csv -- This file is a consolidation of seasons 2014-16 for testing models

mvp-testlabeled.csv -- (see above)

mvp-allblank.csv -- This file is a consolidation of all seasons to predict this 2017 MVP Race

mvp-alllabeled.csv -- (see above)

MLonMVP.ipynb -- This jupyter notebook performs the machine learning on all the csv above

TrendsByEra.csv -- This file has the averages of the MVP candidates every 5 years

NBAMVPStandardsOvertheYears.twbx -- This tableau workbook has visualizations of trends