A Logistic Regression Approach to Predict the NBA Most Valuable Player

Sidhant Rastogi
Class of 2023
Harvey Mudd College
Claremont, CA 91711
srastogi@hmc.edu

Abstract

This paper utilizes a logistic regression classifier model to solve the problem of predicting the Most Valuable Player (MVP) of a National Basketball Association (NBA) season based on player season statistics. Ultimately, the model developed performs well on training and testing data from an NBA player statistics dataset. The paper describes and analyzes its performance on this data. It also identifies some of the drawbacks of the given approach, some potential modifications to the model, and proposes areas for future research in this area and in basketball analytics as a whole.

1 Problem Statement

The problem that we wish to solve is to make a statistical prediction for the Most Valuable Player (MVP) of a given NBA season based on player season statistics–that is, players' totals and percentages across an entire season.

1.1 Background

The Most Valuable Player (MVP) award is a prestigious award given every season in the National Basketball Association (NBA) to the best performing player in the regular season. In basketball, the MVP award is the highest individual accolade that a player can achieve. Every year, some of the best players in the world contend for the award, but only one is given the Maurice Podoloff MVP trophy at the conclusion of the season.

Various factors influence the MVP discussion every season, including individual player statistical success, player roles, and team success. However, the criteria for "best" player varies drastically year-to-year, since the game of basketball is constantly evolving and player skills sets evolve over time. Additionally, there is a degree of subjectivity present due to the nature of the award naming process. The award is voted on by a panel of over 100 sportswriters and broadcasters throughout the United

States and Canada, who have very subjective opinions that are often influenced by narratives that develop about players throughout the season. Ultimately, for these reasons, the task of predicting the MVP of a given season is a difficult task.

This paper takes a statistical approach to the task of predicting the award winner, using machine learning on a dataset of NBA player season statistics to arrive at a prediction of the player with the greatest likelihood of winning the award. Player season statistics were chosen to analyze as opposed to per-game statistics, since total season statistics give a more accurate picture as to the degree of consistent performance over an entire season.

1.2 Previous Research

While there has been some research exists in this area, it is not quite extensive. Specifically, the approach of logistic regression in player ranking and MVP prediction has not been thoroughly examined, which is what this paper seeks to do.

However, some related approaches to these problems do exist. For instance, Peter Li derives some statistical metrics for approaching the question in his blog post:

`https://towardsdatascience.com/nba-mvp-predictor-c700e50e0917`

However, his metrics are largely heuristic-based and not strongly quantitatively supported. Additionally, Jake Levene and Zach Diamandis of the Harvard Sports Analysis Collective used per-game player statistics and linear regression analysis to examine correlations between advanced player statistical metrics and the MVP award in this blog post:

`http://harvardsportsanalysis.org/2018/06/nba-mvp-predictions/`

While they were able to achieve a good accuracy of around 80% correct predictions over the 20 years of data that they tested on, the post remarks very little of the methodology they utilized, and they do not take a learning-based approach to the problem.

So, as we can see, while there are a few related approaches to the problem, there are no scholarly sources that have looked at the problem in considerable depth. Furthermore, as far as has been visible, there are very few machine learning based approaches to the given problem. Basketball analytics is a relatively new and evolving field, so it is reasonable that not much research has been conducted in this particular problem. As more familiarity with machine learning tools enters the analytics community, it is almost certain that more research will be conducted on data-geared approaches to MVP prediction and player ranking.

My analysis builds on the literature by implementing a logistic regression model on a dataset of over 30 variables across over 35 seasons. Making such a predictive model is an important problem, as it can aid NBA players, broadcasters, commentators, and fans alike in making educated predictions.

2 Approach

The approach this paper takes is to implement a logistic regression model for binary classification as either 1) MVP or 0) not MVP. Logistic regression is a statistical model that fits a logistic (sigmoid) curve to given data to model a binary dependent variable. Given a vector of data points x, this can be modeled by the equation

$$p = \frac{1}{1 + e^{-\theta^T x}}$$

where p is the probability that the predicted result is 1. A stochastic gradient descent approach is taken to numerically determine the optimal weight vector θ.

The approach makes extensive use of the data from this Kaggle dataset:

`www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv`

The dataset lists players' season total numbers and percentages from 1950 onwards in over 40 categories, including Points, Rebounds, Assists, Free Throw Percentage, 3 Point Percentage, etc. as well as advanced player metrics such as Player Efficiency Rating (PER), Win Shares, and Box Plus/Minus. The dataset did not have MVP results in it, so that had to be manually inputted as a binary result every season, along with additional data preprocessing. Additionally, only data from 1982 onwards was utilized from the dataset, as years prior contained incomplete data points for many players. This can be attributed to lack of proper data collecting methods early on in the history of the NBA, as well as the introduction of new attributes to the game, such as the introduction of the 3 Point Line in the 1979-80 season. Some redundant and irrelevant variables such as Personal Fouls were also removed from the model, as they either double-counted certain statistics or were irrelevant to the model. Lastly, the model was trained on data from the seasons from 1982-2016 and tested on 2017. In that season, Russell Westbrook was named the MVP of the league.

The Python code for the model can be found at:

`github.com/srastogi1011/MATH-189R/blob/master/mvp_predictor.py`

3 Results

Remarkably, the model performed quite well on the training data, boasting an accuracy of 99.86%. Additionally, on the testing data of the 2017 season, the model was able to derive some suprisingly accurate predictions. It correctly named Russell Westbrook as the predicted MVP of the season. Furthermore, it named James Harden as the 2nd place behind Westbrook, which happened in real life as well. These results are interesting, as that season was a heavily contested race between those two players for the award, with many people believing that the award should have gone to Harden over Westbrook. In that case, this model helps confirm that Westbrook was indeed a deserving recipient of the award over Harden.

The results of the top 10 ranked players, by probability of winning MVP according to the model, is shown in the table below:

Table 1: 2017 Player Rankings by Model

Rank	Name
1	Russell Westbrook
2	James Harden
3	Andre Drummond
4	Draymond Green
5	DeMarcus Cousins
6	Hassan Whiteside
7	DeAndre Jordan
8	Karl-Anthony Towns
9	Nikola Vucevic
10	Nicolas Batum

In general, from the rankings produced by the model, it was clearly visible that better players were ranked higher than worse players, indicating that it performed quite well. Indeed, of the top 20 players that the model produced, 10 of them were named All-Stars, indicating that while the model was decent, there is certainly room for improvement.

3.1 Drawbacks

The dataset itself had some drawbacks that were reflected in the model. For instance, the dataset did not factor in team success throughout the regular season (how many games a given player's team won). This is generally a very relevant factor to the MVP discussion, as good players putting up good numbers on a bad team is not considered as valuable as a good player leading his team to a high win-total. While it does incorporate the variable of win-shares, a metric designed to capture some aspects of team success, this is not a complete picture of team success, as it is not always correlated to total team wins. This may have influenced the model's choice of Russell Westbrook over James Harden in the aforementioned 2017 race, as the main argument for Harden was that he led his team to more wins than Westbrook did that season. Perhaps James Harden really did deserve the award, and running this model on a dataset that factored in more team statistics would reveal that. Also, the dataset is not completely up-to-date, as it only goes up to 2017, missing out on the data of the previous 3 seasons. A more up-to-date dataset might alter the model in terms of which criteria it values more. It's also important to note that while the data provides full summary statistics, historically defensive skills are not well-captured by these statistics, which the NBA itself has admitted. Thus, the model likely under accounts for defense when assessing MVP, and explains why many of the top players it named were mainly offensive-minded players.

Additionally, the model developed had drawbacks that could certainly be improved upon. While the model was able to accurately predict the winner and runner-up for the award in 2017, it was not able to give a very accurate picture of the MVP race that season beyond those two players. One factor that was noticeable was the fact that the model tended to overrate so-called "big men". These are Centers and Power Forwards, some of the largest and tallest players on the court, that typically dominate stats such as Rebounds and Blocks. However, the way that the game is structured,

these players aren't inherently more valuable than smaller players, as they often possess very different skill sets. Of the top 20 players list referenced earlier, 9 out of the 10 players that weren't named All-Stars were big men. A reason for this could be that throughout the early NBA, many big men dominated the league and won a great number of MVPs, which could have skewed the model into overvaluing stats typically dominated by bigger players, such as Field Goal Percentage and Rebounds. However, over time, the league has evolved to value smaller players with skills such as 3 point shooting and playmaking, resulting in the discrepancy.

4 Future Research Proposal

There are many possibilities for future research in this area. The currently existing model can certainly be improved in many ways. First and foremost, the logistic regression approach could be improved somewhat by better feature selection and scaling, as the large number of features may have contributed to some of the variability. Cross-validation could also be performed to improve on generalization error. The problem of not accounting for time-dependence of data could be remedied by testing the model on subsets of the given data, to account for different eras of basketball (such as the 2000s onwards or the 2010s onwards). Additionally, we could take a different approach to model the data. For example, a neural network model might produce a more accurate picture of player rankings. There are also many other interesting questions in the area of basketball analytics that can be answered by this dataset and related datasets. For example, a similar approach could be taken to predict All-Stars or All-NBA team selections.

For the final project, I would like to look mainly at the problem of classifying types of players by skill set based on their statistics. Many players in the NBA are specialized based on factors like position, height, and athletic ability. Some players are excellent 3 point shooters, while others are better passers or defenders. I think player statistics can give a very good idea of the skills they excel at, and as such, can provide a good metric for skill comparison. I want to a used a k-means unsupervised learning approach to classify players into different categories based on these statistics. Additionally, if time permits, I'd like to analyze the correlation of these different categories and player salary data in order to determine if players' contracts are overvalued or undervalued. I'm excited to look into these related problems and build upon the research in this paper.

References

[1] Goldstein, Omri. "NBA Players Stats since 1950." *Kaggle*, 27 Apr. 2018, www.kaggle.comdrgilermonba-players-stats#Seasons_Stats.csv.

[2] Levene, Jake, and Zach Diamandis. "NBA MVP Predictions." *The Harvard Sports Analysis Collective*, 19 June 2018, harvardsportsanalysis.org201806nbamvp-predictions.

[3] Li, Peter. "NBA MVP Prediction Model." *Medium*, Towards Data Science, 20 Apr. 2019, towardsdatascience.comnbamvppredictorc700e50e0917.